Image Quality Assessment for Population Cardiac MRI: From Detection to Synthesis

Le Zhang

Department of Electronic and Electrical Engineering

University of Sheffield

A thesis submitted for the degree of

Doctor of Philosophy

March 2019

This thesis is dedicated to
someone
for some special reason

Acknowledgements

First and foremost I wish to thank my supervisor Prof. Alejandro Frangi, for giving me the opportunity to pursue a PhD degree and work alongside him. He has been a constant source of support and inspiration for me, and it has been a great pleasure and privilege to carry out my PhD research under his supervision. I still do remember my first PhD meeting with him, when he suggested me to do lots of reading and particularly machine learning, which has probably influenced my research career the most.

Dozens of people have helped and taught me immensely at the CISTIB group over the last four years. I am very thankful to everyone, especially to Bo, Yawen, Mohsen and Hamid for all the good and fun memories. Special thanks to Dr. Marco Pereañez for his help and guidance, without whom I would not have been able to complete the research work presented in this thesis.

I also need to thank to Prof. Steffen Petersen from Queen Mary University of London, Prof. Stefan K. Piechnik, and Prof. Stefan Neubauer from Univeristy of Oxford for their clinical feedback and for providing almost all the CMR imaging data presented in this thesis.

Finally, I am deeply grateful to my parents and sister for their constant support and love, who have given me all the strength and courage to follow my dreams.

Abstract

Cardiac magnetic resonance (CMR) images play a growing role in diagnostic imaging of cardiovascular diseases. Left Ventricular (LV) cardiac anatomy and function are widely used for diagnosis and monitoring disease progression in cardiology and to assess the patient's response to cardiac surgery and interventional procedures. For population imaging studies, CMR is arguably the most comprehensive imaging modality for non-invasive and non-ionising imaging of the heart and great vessels and, hence, most suited for population imaging cohorts. Due to insufficient radiographer's experience in planning a scan, natural cardiac muscle contraction, breathing motion, and imperfect triggering, CMR can display incomplete LV coverage, which hampers quantitative LV characterization and diagnostic accuracy.

To tackle this limitation and enhance the accuracy and robustness of the automated cardiac volume and functional assessment, this thesis focuses on the development and application of state-of-the-art deep learning (DL) techniques in cardiac imaging. Specifically, we propose new image feature representation types that are learnt with DL models and aimed at highlighting the CMR image quality cross-dataset. These representations are also intended to estimate the CMR image quality for better interpretation and analysis. Moreover, we investigate how quantitative analysis can benefit when these learnt image representations are used in image synthesis.

Specifically, a 3D fisher discriminative representation is introduced to identify CMR image quality in the UK Biobank cardiac data. Additionally, a novel adversarial learning (AL) framework is introduced for the cross-dataset CMR image quality assessment and we show that the common representations learnt by AL can be useful and informative for cross-dataset CMR image analysis. Moreover, we utilize the dataset invariance (DI) representations for CMR volumes interpolation by introducing a novel generative adversarial nets (GANs) based image synthesis framework, which enhance the CMR image quality cross-dataset.

Contents

Acronyms

AD Alzheimers disease

AIQA Automatic Image Quality Assessment

AL adversarial learning

AI artificial intelligence

AV atrioventricular

BPTT Backpropagation Through Time

CMR cardiac magnetic resonance

CMRI Cardiac Magnetic Resonance Imaging

CNN convolutional neural networks

CO Cardiac Output

CSAE convolutional sparse autocoder

CT computed tomography

CVD cardiovascular disease

DA Dataset Adaptation

DL deep learning

ECG electrocardiogram

ED end-diastole

EF Ejection fraction

ES end-systole

FOV field-of-view

GAN Generative adversarial network

GE gradient echo

LA left atrium

LAX long-axis

LEG late gadolinium enhancement

LSTM long short term memory

LV left ventricle

MAS missing apical slice

MBS missing basal slice

MLMT multi-label multi-task

MR Magnetic resonance

MRF Markov random field

MRI Magnetic resonance imaging

MS multiple sclerosis

MV mitral valve

PET positron emission tomography

PI privileged information

RBM Restricted Boltzmann Machine

RF radio frequency

RNN Recurrent Neural Network

RV right ventricle

SA short axis

SAE Stacked Autoencoder

SAX short-axis

SE Spin echo

SSAE stacked sparsely autoencoder

SSFP cine steady-state free precession

SV stroke volume

UKBB UK Biobank

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Why Population Imaging?

Images represent complex object mappings, and when digitized, they can be as multidimensional data containing discrete image pixels. Such images, when generated for specific purposes using suitable data processing systems to elucidate physical factors and human tissues, do not suffer from the limitations that are present in the previously used clinical diagnostic methods such as auscultation and touching.

Population imaging has become an indispensable component of disease prediction, treatment, and risk prevention and is becoming increasingly important. Magnetic resonance imaging (MRI), positron emission tomography (PET), computed tomography (CT), cone beam CT, and ultrasound imaging are the most commonly used population imaging technologies in clinical examination, diagnosis, treatment, and decision making. These technologies are widely used in medical practice because they enable visualization of body without inflicting trauma and are evolving into a unique branch of medical science.

In today's world of big data, we are rapidly accumulating various types of population imaging data. In this section, we consider several characteristics of population imaging data and describe why these data are important in both engineering and clinical analysis.

1.1.1 High Number of Dimensions

Due to the influence of imaging sources, modalities, parameters, imaging times, and other factors, population imaging data are complex, diverse, and context-rich [270]. Such data involve many variables that reflect different aspects of the same object, *i.e.*, they are high-dimensional. Detailed descriptions of objects can be obtained given the characteristics of various aspects of the objects and the relationships among these characteristics. Consequently, many variables are generated to create abstract descriptions of complex objects,

and vector data are created to form an abstract high-dimensional data space, which is the basis of the descriptions of various characteristics of the objects and the interrelationships among these characteristics.

1.1.2 Data Mining

Population imaging can reflect the most basic and intuitive information about the human body. For example, population imaging can provide basic and intuitive information about the heart, brain, and liver at the organ level, as well as about the cardiovascular and nervous systems at system and cell levels. The relationship between different types of diseases, diseases and occupations, and blood groups and other factors can be extracted from massive population imaging data. For example, based on texture features, important association rules can be determined from massive image data sets, and these rules can be applied to identify hidden information in mammograms [52].

1.1.3 Data Fusion

Although many clinical diagnoses are based on single modality imaging, the appearance and description of single modality feature information on population imaging are insufficient to be the basis of accurate diagnosis in case of specific diseases, such as multiple sclerosis (MS) lesion detection and segmentation. Combining population imaging data from different sources, modalities, and representations, *i.e.*, image fusion, is often required to synthesize image features from different imaging mechanisms as well as to preserve and strengthen respective feature information, and to display an image to obtain a more accurate description of the object [130]. Image fusion facilities a comprehensive analysis and extraction of target features. For example, fusion results of MRI and PET images of the brain can provide information about soft tissue structure (MRI) and metabolism (PET) [83]. In tumor localization tasks, PET/CT can help accurately locate lesions and can provide tissue and cellular metabolic information [186].

Population imaging has broad application prospects in disease diagnosis and treatment. Currently, for many diseases, population imaging has been helping clinicians achieve more accurate diagnosis. In future, it is expected that independent population imaging diagnostic systems will be used to provide reliable evaluations of pathological sections, which will reduce the cost of human resources in hospitals. In addition, the quality and efficiency of pathological diagnosis will improve. The improvement and practicality of population imaging diagnosis will help achieve standardization and quality diagnoses under different conditions.

Although population imaging diagnosis has broad application prospects, it can only be applied to a narrow spectrum of diseases, and its accuracy needs to be improved. At present, problems associated with population imaging diagnosis are primarily attributed to the difficulty involved in acquiring sufficient training data. Population imaging data have the characteristics of single access and condition; thus, such data have high intellectual property and economic value. Integrating resources, solving the problems associated with intellectual property rights, and obtaining large amount of data required by artificial intelligence technology are key issues faced by population imaging diagnosis technology research. Solving these problems will allow population imaging diagnosis technology to move rapidly from laboratory to clinical environments, thereby benefiting an increasing number of patients.

1.2 Clinical Background and Motivation

In this section, we briefly introduce the anatomy and physiology of the human heart, with a particular focus on electromechanical events that occur during the cardiac cycle. In addition, we discuss MRI, highlighting the specific techniques that are used particularly in cardiac MRI, such as electrocardiogram (ECG) gating, respiratory motion compensation, and cine imaging. We briefly summarize the criteria used to evaluate the cardiac magnetic resonance (CMR) image quality, and describe a numerical score that is obtained for each type of imaging sequence to show the image quality of its modules and the image quality of the overall CMR study. Specifically, we describe the criteria used to evaluate the quality of cine steady-state free precession (SSFP) images and the quantitative indices used in cardiac function analysis.

1.2.1 Anatomy and Function of the Heart

The human heart pumps blood to the entire body via the circulatory system, which provides oxygen and nutrition to tissues and removes carbon dioxide and other wastes [223]. The human heart has four chambers, *i.e.*, two atria (upper chambers) and two ventricles (lower chambers), as shown in Figure 1.1. The antrioventricular septum separates the two sides of the heart. Unless there is a septal defect, the two sides do not interact directly; however, they function together.

There are two separate circulatory pathways through the heart [175]: *i.e.*, the pulmonary circuit and the systemic circuit. In the pulmonary circuit, deoxygenated blood goes to the right side of the heart, *i.e.*, the right ventricle (RV). From there, the blood goes to the lungs and the oxygen is absorbed. Finally, the oxygenated blood returns to the left atrium (LA)

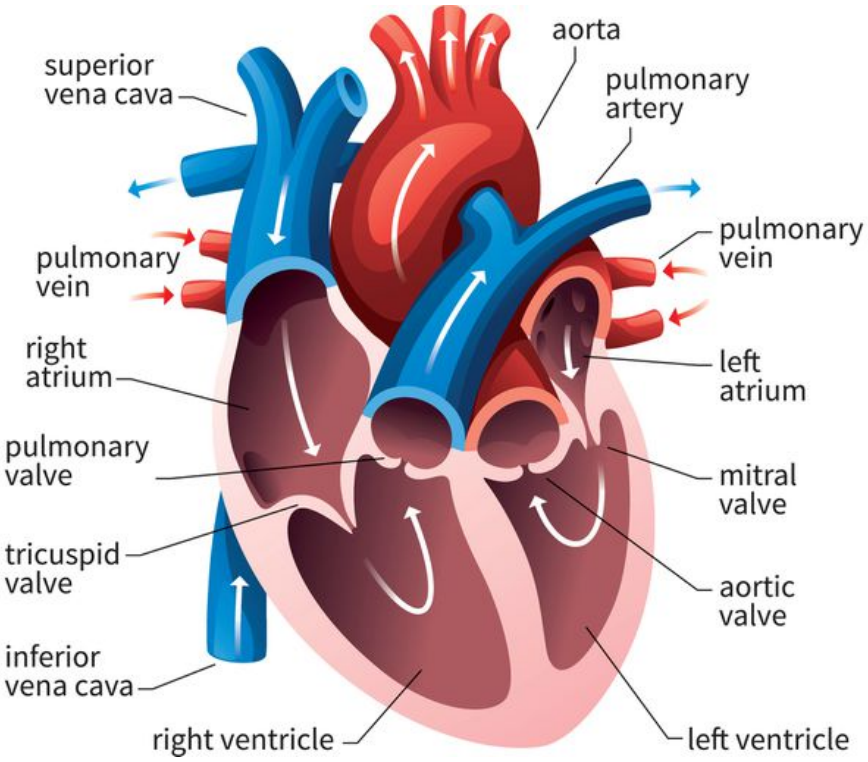


Figure 1.1: Anatomy of the human heart (blood vessels, ventricles and atria). The arrows show the flow of blood [1].

through the pulmonary vein. In the second pathway, the systemic circuit, the left ventricle (LV) pumps the oxygenated blood into the aorta and arterial circulation. The atria and the ventricles are separated by the bicuspid (mitral) and tricuspid atrioventricular (AV) valves. The papillary muscles are projections of the ventricular muscles and they attach to the cusps of the AV valves (Figure 1.2). There are four stages in a healthy heart contraction. First, the heart is relaxed in early diastole. Second, in the atrial systole stage, the atrium contracts to push blood into the ventricles. Third, the ventricles keep the volume constant and contract until the ventricles are empty. In the last stage, they stop contracting, relax, and repeat the loop. Valves maintain blood flow in one direction and prevent its backflow.

The ventricular wall comprises three layers: the outermost (epicardium), middle (myocardium), and inner (endocardium) layers; the myocardium contains the muscle cells [77]. As seen in Figure 1.2, the ventricular wall primarily comprises muscle cells and the LV wall is thicker than the RV wall, because sufficient blood pressure is required to pump oxygen-containing blood to different parts of the body. [167].

[1]The figure has been adapted from: https://www.thoughtco.com/evolution-of-the-human-heart-1224781, accessed on 5 January, 2019

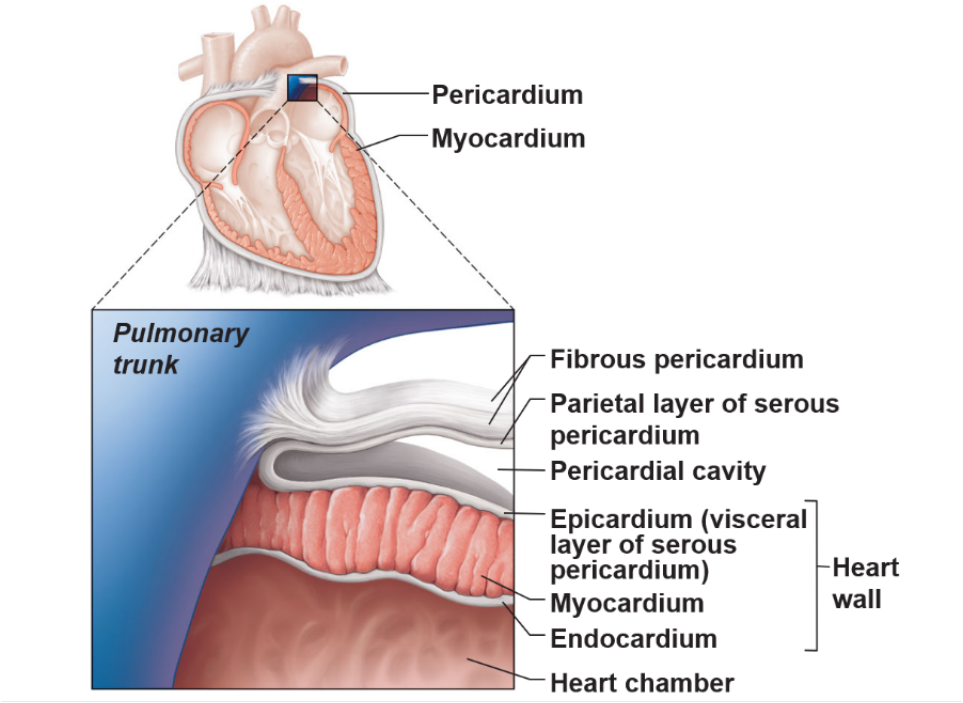


Figure 1.2: Layers of tissue that comprise the heart wall: 1. epicardium 2. myocardium 3. endocardium[2].

1.2.2 Cardiac Magnetic Resonance Imaging

MRI is a noninvasive and nonionizing imaging technique that can produce tomographic images with unmatched soft tissue contrast. Unlike X-ray and CT, MRI is based on the absorption and emission of energy in the radio frequency (RF) range of the electromagnetic spectrum rather than on ionizing radiation [43]. MR images are produced based on the frequency of the RF energy being absorbed and emitted by the anatomical tissues and the spatial variations in the phase. Under the influence of a strong magnetic field, magnetically aligned hydrogen nuclei generate transverse magnetization as a response to applied RF pulse sequences, which is captured by the scanner and then reconstructed as an image. The magnetic response of the tissues across time is shaped by atomic properties that differ depending on tissue type [19].

Cardiac MRI produces detailed images of the heart's interior and surrounding structures using powerful magnetic fields, radio waves and computers. Cardiac MRI is used to detect or monitor heart diseases and to assess cardiac anatomy and function in patients with congenital and postnatal heart diseases [227]. In some cases, cardiac MRI can provide the

[2]The figure has been adapted from: https://www.easynotecards.com/notecard-set/89049, accessed on 8 January, 2019

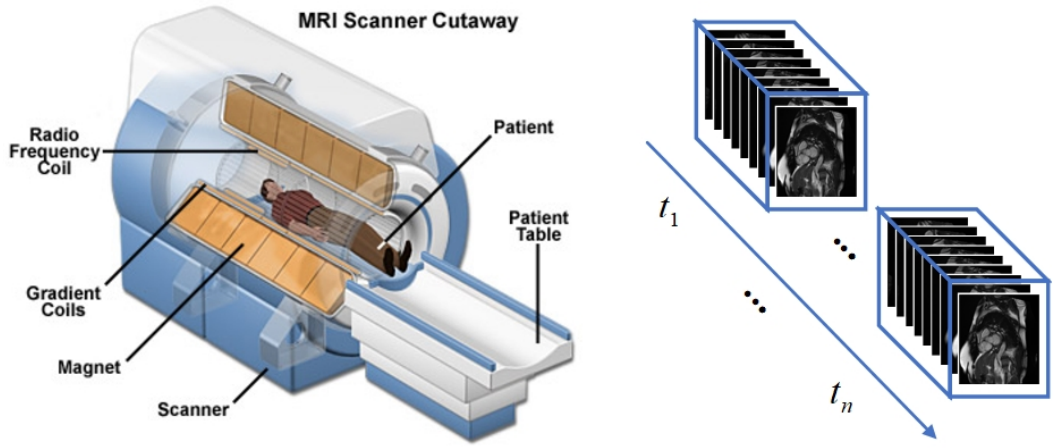


Figure 1.3: MRI Scanner and cardiac MRI time series acquisition[3].

best image of the heart. In the past, imaging difficulties associated with cardiac and respiratory motion and long acquisition time have limited the use of technology. However, with recent advancements in spatial and temporal resolution and reduced scanning times, CMR imaging has been increasingly applied to diagnose cardiovascular disease (CVD) [25]. In particular, parallel imaging is commonly used to reduce imaging time [150], and ECG gating is often required to reduce cardiac motion artifacts.

Cine cardiac MRI: The temporal dynamics of the heart chambers can be visualized using cine MRI, which can be also used for functional assessment. Data points for each cardiac phase are acquired at multiple time points by filling separate k-space lines over several cardiac cycles, which results in the reconstruction of an image for all cardiac phases. The images can be viewed as a movie sequence. Breath-holding cine MRI allows acquisition of k-space data in segments for each cardiac phase. Thus, acquisition times can be further reduced albeit at the expense of reduced temporal resolution, which can be circumvented with echo-sharing [64].

As a pulse sequence, spoiled gradient echo (GE) technique has been preferred for functional imaging because it requires very short repetition times. With the GE sequence, the blood-pool (BP) appears bright and the contrast between the endocardium and BP makes it suitable for ventricular function assessment and analysis [193]. An alternative to the GE approach, SSFP [106] enables relative independence of contrast from blood flow and high speed acquisition, which has been a limitation due to very short RF repetition times [24].

[3]The figure has been adapted from: https://devblogs.nvidia.com/nvidia-digits-alzheimers-disease-prediction/, accessed on 9 January, 2019.

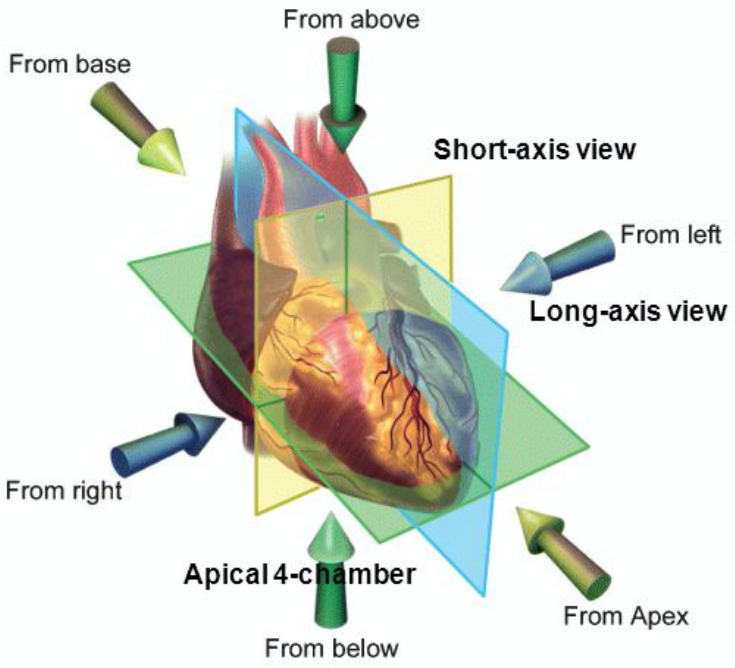


Figure 1.4: Three standard cardiac cine imaging planes [78]: vertical and horizontal long-axis planes, (*i.e.*, two-chamber and four-chamber views respectively[4]), and the short-axis mid-ventricle plane.

On the other hand, a spin echo (SE) sequence is more suitable for anatomical imaging. An SE sequence requires longer acquisition times; however, relatively few metal artifacts are introduced. SE plays a secondary role in CMR imaging; however, it is used for some procedures, such as structural assessment of ventricular abnormalities [64].

Cardiac MRI planes: Three-dimensional (3D) high-resolution cine imaging of the heart is challenging due to long acquisition times, cardiac motion, and repeated breath holds. Therefore, stacks of thick two-dimensional (2D) slices are acquired from different imaging planes for multiple cardiac phases. Standard cardiac cine imaging planes (Figure 1.4) include one short-axis (SAX) plane and two long-axis (LAX) planes (four and two chamber views, respectively) [78]. To identify these planes during acquisition, first scout imaging is performed with a fast single-shot sequence; then, the LAX planes are identified along a line extending from the cardiac apex to the center of the mitral valve (MV). Finally, the SAX plane is determined. The SAX plane extends horizontally perpendicular to the LAX of the heart in the middle of the LV. Since ventricular volume measurements

[4]The figure has been adapted from: https://slideplayer.com/slide/8700066/, accessed on 11 January, 2019.

produce cross-sectional sections that are almost perpendicular to the myocardial boundary, typically, SAX stacks are used to measure ventricular volume. In this way, partial volume effects can be reduced and ventricular measurement accuracy can be improved [88]. Typically, the in-plane and through plane resolutions of the SAX stacks are 1-2.5 mm and 8-10 mm respectively.

1.2.3 Criteria for CMR Image Quality Evaluation

To assess cardiovascular pathologies, ventricular volume and mass must be quantified. Note that CMR is the stanard reference imaging technique for quantitative analysis [235][126]. Cine SSFP, late gadolinium enhancement (LEG) images, and first-pass stress perfusion images are commonly used CMR image modules for medical image analysis. However, cine CMR image quality is limited by known common image artifacts and other factors and, image quality evaluation criteria were first defined a priori. Each type of image sequences, such as the CMR image modules mentioned above, can be evaluated using specific criteria that return a numerical score that indicates the image quality of the overall CMR study and its three modules. In a study [114], a numeric scoring system ($0\sim3$) was used to assess 35 qualitative criteria. In this system, a higher number indicates a poorer image quality (as shown in Figure. 1.5). Among these qualitative criteria, twelve criteria were used to evaluate the quality of cine SSFP images, specifically, the stack of SAX cine images was assessed using criteria $1\sim11$ and the criteria 12 refer to the LAX cine image.

LV coverage: The first quality criterion assesses the stack of SAX cine images, *i.e.*, the LV coverage. To measure the cardiac volume and function accurately, the full LV coverage from the basal slice to apical slice is required. If the basal slice is missing, atrial chamber is not visible in end-systole (ES); consequently, it cannot be ensured that the heart is fully covered from base. A missing apical slice, which is defined as the LV cavity still visible at ES, is another frequent realistic limitation regarding the LV coverage. A missing basal slice will have more significant impact on the cardiac volume calculation; thus, the image quality score will be higher for a missing basal slice than for a missing apical slice. Except for basal and apical slices, a missing mid-ventricular slice will also result in a penalty. Specifically, a missing basal slice (or when $\geqslant1$ additional slice(s) missing) is assigned a score of 3, and a missing basal slice is given a score of 2. However, to ensure that the influence of this criterion is balanced by other criteria, the maximum score for this criterion is limited to five. Regarding the remaining quality criteria, i.e., criteria from 2 to 7, which include wrap around artifacts, respiratory ghost, cardiac ghost, image blurring/mis-triggering, metallic artifacts, and shimming artifacts, a single SA slice is assigned a score of 1 if the artifact impedes the visualization of more than one-third of the LV endocardial border at ES and/or

Qualitative Criteria					
LV-Function cine SSFP	0	1	2	3	Maximum Score
1. LV coverage	Full coverage	-	Apex not covered	Base or >=1 slice in the stack missing	5
2. Wrap around	No	1 slice	2 slices	>=3 slices	3
3. Respiratory ghost	No	1 slice	2 slices	>=3 slices	3
4. Cardiac ghost	No	1 slice	2 slices	>=3 slices	3
5. Image bluring/mis-triggering	No	1 slice	2 slices	>=3 slices	3
6.Metallic artifacts	No	1 slice	2 slices	>=3 slices	3
7. Shimming artifacts	No	1 slice	2 slices	>=3 slices	3
8. Signal loss (coil inactive)	Actived	-	Not actived	-	2
9. Orientation of stack	correct	-	incorrect	-	2
10. Slice thickness	<= 10 mm	11-15 mm	-	>15 mm	3
11.Gap	< 3 mm	3-4 mm	-	> 4 mm	3
12. Correct LV long axes	>= 2	1	-	None	3
LV function score					21
Late Gadolinium Enhancement	0	1	2	3	
13. LV coverage	Full coverage	-	Apex not covered	Base or >=1 slice in the stack missing	5
14. Wrap around	No	1 slice	2 slices	>= 3 slices	3
15. Respiratory ghost	No	1 slice	2 slices	>= 3 slices	3
16. Cardiac ghost	No	1 slice	2 slices	>= 3 slices	3
17. Image bluring/mis-triggering	No	1 slice	2 slices	>= 3 slices	3
18.Metallic artifacts	No	1 slice	2 slices	>= 3 slices	3
19. Signal loss (coil inactive)	Actived	-	Not actived	-	2
20. Slice thickness	>= 10 mm	11-15 mm	-	> 15 mm	3
21.Gap	< 3 mm	3-4 mm	-	> 4 mm	3
22. Correct LV long axes	>= 2	1	-	None	3
LGE Score					19
First-Pass Perfusion	0	1	2	3	
23. LV coverage	>= 3 slices	-	2 slices	1 slice	3
24. In-plan spatial resolution	< 3 mm	-	-	>= 3mm	3
25. Acquisition window	< 150 ms	-	150-250 ms	> 250 ms	3
26. Patient preparation	Drugs + caffeine stopped	Drugs not stopped	Caffeine not stopped	Drugs + caffeine not stopped	3
27. Wrap around	No	1 slice	2 slices	>= 3 slices	3
28. Respiratory ghost	No	1 slice	2 slices	>= 3 slices	3
29. Cardiac ghost	No	1 slice	2 slices	>= 3 slices	3
30. Image bluring	No	1 slice	2 slices	>= 3 slices	3
31. Metallic artifacts	No	1 slice	2 slices	>= 3 slices	3
32. Singal loss (coil inactive)	Actived	-	Not actived	-	2
33. Breathing motion	-	Drift	-	Abrupt	3
34. Mis-triggering	None	1-2 mis-triggers	-	> 2 mis-triggers	3
35. Rhythm	Sinus	-	-	Atrial fibrillation	3
Perfusion Score					20
Total Qualitative Score					60
Quantitative Criteria					
36. LGE SNR: anterior wall	> 5	2-5	< 2	-	2
inferior wall	> 5	2-5	< 2	-	
37. First Pass %SI increase: anterior wall	> 200%	100-200%	< 100%	-	2
inferior wa	> 200%	100-200%	< 100%	-	
Total Quantitative Score					4
Global Quality Score					64

Figure 1.5: CMR image quality definition (adapted from [114]). The total qualitative score is the sum of the SSFP, LGE and perfusion images scores.

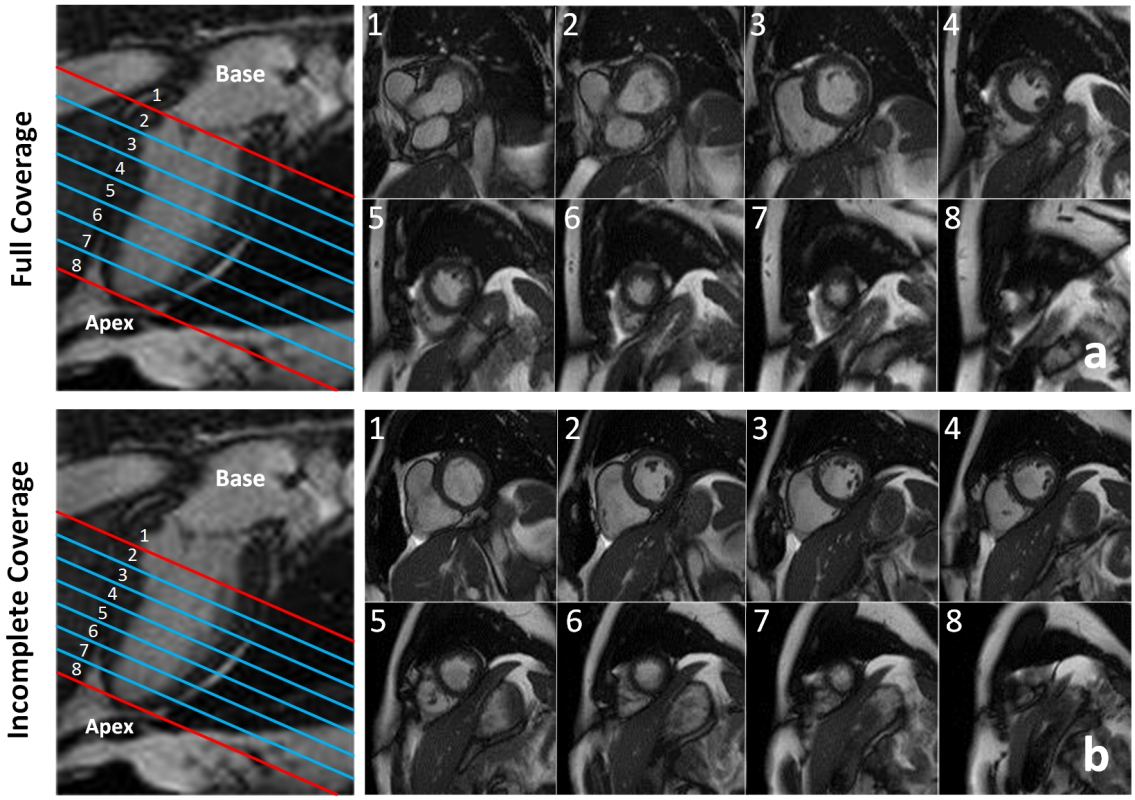


Figure 1.6: *Top:* A typical two-chamber view cardiac MRI with eight slices fully covered from base to apex and SAX view volume with whole coverage (slice 1 is the basal slice); *Bottom:* A typical two-chamber view cardiac MRI with eight slices incompletely covered from base to apex and SAX view volume with missing basal slice (slice 1 is not the basal slice). In each rectangle, from top to bottom, rows correspond to adjacent axial slices.

ED. Scores of 2 or 3 scores are assigned if the artifact involved 2 slices or $\geqslant$3 slices. In this assessment, the quality of RV coverage was not measured (see Figure 1.6).

Wrap around artifacts: A wrap-around artifact (criterion 2) is one of the most common MR artifacts. Such artifacts are usually recognized as anatomic parts that intrude into the area of interest [114], such as an object whose dimensions exceed the defined field-of-view (FOV) (Figure 1.7).

Respiratory and cardiac ghosts: Typically, in clinical MRI, ghost artifacts occur due to patient-related causes, such as cardiac and respiratory motion (criteria 3 and 4), and they usually occur during image acquisition. For example, data sampling and reconstruction causes a mis-mapping of the signal when a spin moves during the time of excitation. When the amplitude of the periodic motion or the signal intensity of the moving tissue increases,

Figure. 1.7 $\sim$ Figure. 1.10 are adapted from [114].

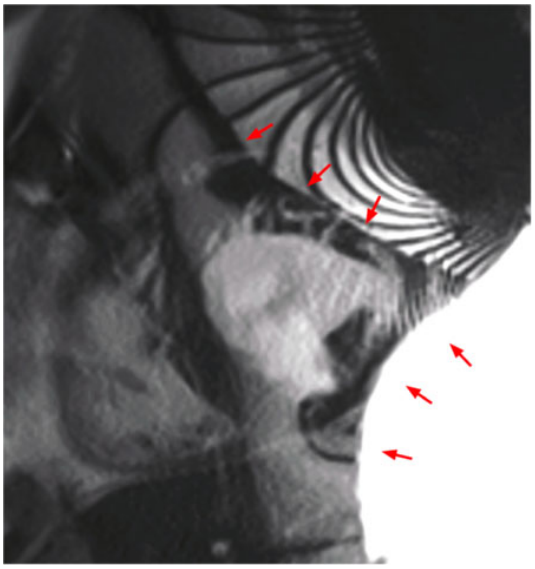


Figure 1.7: Wrap-around in a cine SSFP sequence. The chest wall, which is outside the FOV, protrudes into the LV (shown as red arrows).

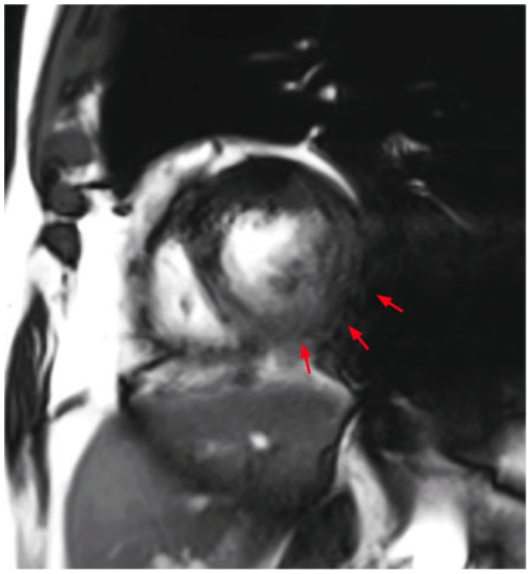


Figure 1.8: Image blurring or mis-triggering in a cine SSFP sequence. Blurred aspects are indicated by red arrows.

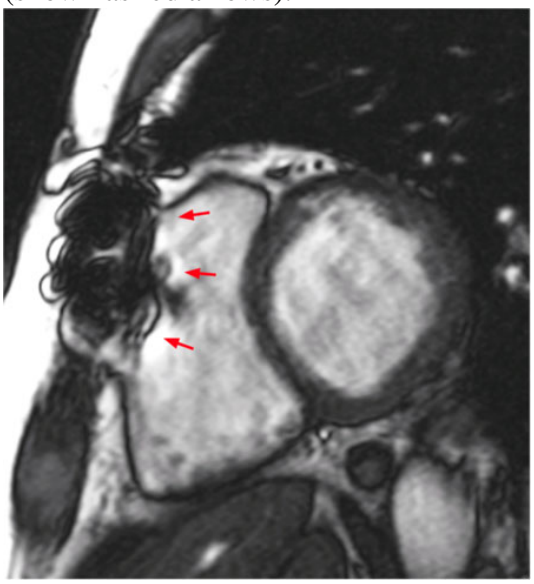


Figure 1.9: Metal artifact in a cine SSFP sequence. Ferromagnetic material disturbs the magnetic field locally.

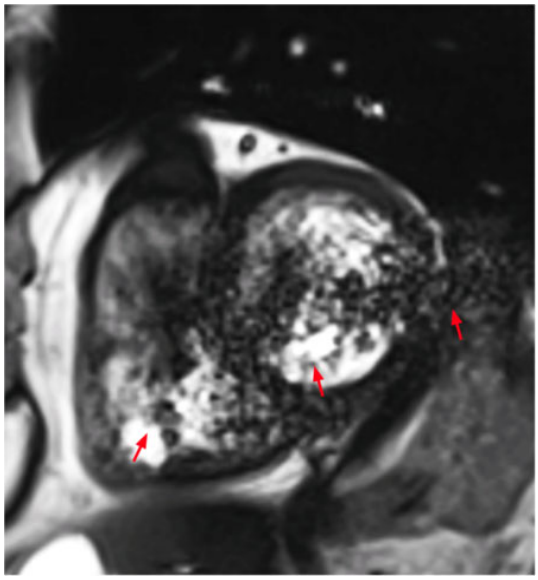


Figure 1.10: Shimming artifact in a cine SSFP sequence. Magnetic field inhomogeneities produce a dark band and flow-related artifacts on the LV (red arrows).

the intensity of these ghost artifacts will increase. Diffuse image noise will be generated and propagate widely along the phase-encode direction. However, discrete ghosts will not be formed if the motion is nonperiodic (e.g., peristalsis).

Image blurring or mis-triggering: This artifact (criterion 5) is primarily caused by extrasystoles, irregular heartbeats related to atrial fibrillation, mis-triggering of the R-wave, or respiratory motion. Regarding the standard for retrospectively gated cine imaging, blurring in SSFP images generally occurs because signals that are collected in the different phases of the cardiac cycle are used to reconstruct a specific phase of the cardiac cycle (see Figure 1.8).

Metal artifacts: Metal (primarily iron) can deflect the magnetic field, which causes metal artifacts (criterion 6) and changes the resonance frequency beyond a given range. Consequently, protons do not react properly to the excitation pulse, *i.e.*, they are not properly excited, resulting in signal degradation/image distortion (Figure 1.9).

Shimming artifacts: Shimming artifacts (criterion 7) are caused by the inhomogeneity of the main magnetic field, which particularly influences SSFP acquisition schemes and may lead to band artifacts (dark bands on images caused by non-frequency sound) and/or flow-related artifacts (Figure 1.10).

1.2.4 Indices of Cardiac Function

Quantitative analysis of the cardiac ventricles using imaging data begins with the delineation of the endocardial and epicardial boundaries of the myocardium. Once contours are defined for each slice in the stack of images, local and global volumetric measurements can be performed to assess ventricular function and mass [88]. These measurements have clinical importance in the diagnosis of cardiac pathologies, such as cardiac hypertrophy and dilated cardiomyopathy [192]. Myocardial mass (M) is a particular example. Here, M corresponds to the weight of the heart muscle and is calculated by multiplying the ventricular volume (V) calculated from the contours by the density of the myocardium ($\rho_m = 1.05 g/cm^3$) [66]. LV myocardial mass is calculated as: $M_{LV} = V_{LV} \cdot \rho_m$.

Global functional indices indicate the overall ability of cardiac ventricles to supply blood to the rest of the body [68]. They require myocardial contouring at at least two points in the cardiac cycle: end-diastole (ED) and end-systole (ES). As a global functional index, the stroke volume (SV) corresponds to the volume of oxygenated blood pumped from the LV in each cardiac cycle, which is equal to the difference between the LV volumetric measurements at ED and ES phases: $SV = V_{ED} - V_{ES}$. The ejection fraction (EF) is the fraction of SV (ejected blood) with respect to the volume of the filled heart (V_{ED}) and is defined as $EF = SV/V_{ED}$. Cardiac output (CO) is a functional index that is defined as the

amount of blood ejected from the LV per minute and is equal to the *SV* multiplied by the heart rate. Although global indices are good indicators of functional abnormalities, they do not convey specific information regarding the parts of the ventricle that have a reduced or altered contractile function. In addition, there may be instances where global measurements fall within the healthy range while the wall motion is abnormal. Therefore, local functional analysis of the ventricular wall, including wall thickening and strain analysis, is performed. Such analyses can precisely identify reversibly injured yet viable parts of the myocardium [26]. A survey study [68] on local cardiac wall motion provides more detailed information.

1.3 Challenges in CMR Image Quality Assessment

For population imaging studies, CMR imaging provides a noninvasive access to cardiac anatomy and function [182]. Quantification of ventricular anatomy and function from large population imaging studies or patient cohorts from extensive clinical trials is vital to assess cardiovascular pathologies. Such quantification requires automatic image quality assessment and image analysis tools. The technical limitations of imaging systems necessitate designing of robust and accurate image analysis frameworks for quantitative assessment. In addition, accurate predictive performance, ease of use, and interpretability are essential if such frameworks are to be used in clinical diagnostics.

Limited imaging artifacts in CMR imaging. Few guidelines, clinical or otherwise, objectively establish what constitutes a good medical image and a good CMR study [242]. To ensure consistent quantification of CMR data, automatic assessment of complete LV coverage is the first step. LV coverage is still assessed by manual visual inspection of CMR image sequences. However, manual assessments are subjective, repetitive, error prone, and time consuming [9]. An automatic coverage assessment that can intervene promptly and adjust the data acquisition process, and/or discard images with incomplete LV coverage is required. Analysis of a set of images that includes images with incomplete LV coverage would return inaccurate aggregated statistics for a cohort. The most common causes of incomplete LV coverage are lack of basal slices (no atrial chamber visible in ES, thus no certainty that the base of the heart is covered completely) and lack of an apical slice (LV cavity still visible at ES) [114]. Technological advances in MRI hardware and pulse sequencing have helped achieve faster CMR acquisition, full heart imaging, and motion compensation; however, certain challenges remain. For example, the CMR protocol of the UK Biobank (UKBB) flags 4% of all CMR examinations as unreliable or non-analyzable image data due to incomplete heart coverage [21]. While 4% may be a small proportion, the challenge is to automatically sift through the entire database to identify and exclude

these cases from further quantitative analysis. Methods for objective detection of basal and apical imaging planes are relevant because absence affects diagnostic accuracy as well as anatomical and functional LV quantification.

Limited quality assessment methods in video processing. In video processing, automatic image quality assessment represents a well-developed corpus of techniques to detect image distortions that commonly occur in multimedia communication [202] [264]. However, these distortions generally differ significantly from those that occur in medical imagery. No-reference based image quality assessment [87], [164] is relevant for medical imaging data because it is not possible to collect data that does not contain artifacts or have some level of image degradation. Typically, in practice, only images with incomplete LV coverage are available as input to CMR image processing applications. While assessment methods attempt to compare an available image to a hypothetical high-quality image [110], the final image quality is estimated based solely on the characteristics of the assessed image.

Idenfitication of correspondences in cross-modality imaging data. In medical image analysis, for image quality assessment purposes, it is sometimes convenient or necessary to infer an image in one modality from an image in another modality. A significant challenge for CMR slice pose estimation comes from differences between data sources. These differences involve tissue appearance and/or the spatial resolution of images that are sourced based on different physical acquisition principles or parameters. Such differences make it difficult to generalize algorithms trained on specific datasets to other data sources. This is problematic when the source and target datasets differ and even more problematic when the target dataset contains no labels. Under such conditions, it is highly desirable to learn a discriminative classifier or other predictor in the presence of a shift between training and test distributions, which is called *dataset invariance*. Various approaches to achieve dataset adaptation have been explored under many facets. Among the existing cross-dataset learning studies, dataset adaptation has been adopted for re-identification based on the expectation that labeled data from a source dataset can provide transferable identity-discriminative information to a target dataset. A previous study [97] explored the possibility of generating multi-modal images from single-modality imagery. Other existing studies [134] [151] have employed multi-task metric learning models to benefit the target task. However, these studies primarily focus on linear assumptions.

1.4 Thesis Contributions

The following sections present the primary contributions of this thesis to the analysis of cardiac cross-dataset imaging data. In addition, how the aforementioned challenges can be addressed is considered.

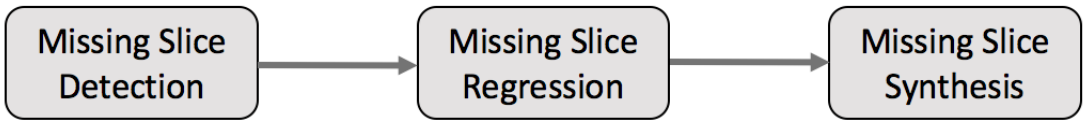


Figure 1.11: Primary contributions (from the CMR image quality detection to image quality recovery).

Full LV Coverage Assessment in CMR imaging with Fisher Discriminative 3D CNN.

Full LV coverage is a basic criterion for CMR image quality. Complete LV coverage from base to apex is required to assess functionality and to measure cardiac volume accurately. Incomplete LV coverage is identified manually through visual inspection, which is time consuming and is usually done retrospectively when assessing large imaging cohorts. In this regard, we propose a novel automatic method to determine LV coverage from CMR images using Fisher-discriminative 3D (FD3D) convolutional neural networks (CNN). Compared with our previously proposed method [271] that used 2D CNN, this approach utilizes spatial contextual information in CMR volumes, extracts more representative high-level features, and enhances the discriminative capability of the baseline 2D CNN learning framework, thereby achieving superior detection accuracy. A two-stage framework is proposed to identify missing basal and apical slices in CMR volume measurements. First, the FD3D CNN extracts high-level features from the CMR stacks. Then, these image representations are used to detect missing basal and apical slices. Compared with the traditional 3D CNN strategy, the proposed FD3D CNN minimizes within-class scatter and maximizes between-class scatter. We constructed a large dataset with more than 5,000 independent volumetric CMR scans and performed extensive experiments to validate the proposed method. The proposed approach outperformed previous methods that rely on 2D CNN. In addition, the proposed method can be adapted for LV coverage assessment of other types of CMR image data. After the quality control for missing slice detection, we can classify the images into good/bad image quality. However, we still do not know the position of each slice, especially the missing slice position. We continue our work from missing slice detection to the next step - regression for slice pose (position and orientation).

Automatic Plane Pose Estimation Across Cardiac Cine MRI Datasets via Deep Adversarial Ranking Nets with Privileged Information.

Cardiac function parameters, such as the EF and CO of both ventricles, are the most immediate indicators of normal/abnormal cardiac function. To compute these parameters, accurate measurement of ventricular volumes at ED and ES is required. Accurate volume measurements depend on the correct identification of the ventricle pose, particularly the positions and orientations, in CMR sequences that provide full LV and RV coverage. This thesis proposes an adversarial learning (AL) CNN-based approach that detects and localizes the CMR slices in an image volume independent of image acquisition related idiosyncrasies, such as the imaging device, magnetic field strength and variations in protocol execution. Furthermore, we incorporate additional information, such as cross-view information, into the training phrase. Note that such information has been referred to as privileged information (PI). The proposed model is trained on multiple cohorts of different provenance and unified using PI loss with different MRI viewing planes to learn the appearance and localize the short-axis view planes of the heart. To the best of our knowledge, this is the first study to tackle fully-automatic detection and pose localization of bio-ventricular slices in CMR volumes in a dataset-invariant manner. We achieve this by maximizing the ability of a CNN to ordinarily regress the positions and orientations of short-axis view planes within a single dataset, while minimizing the ability of a classifier to discriminate image features between different data sources. The regression parameters are important since they provided the information for people to understand the 'where' and 'how' the sub-optimal image quality is. With the development of generative adversarial models, we can use these information to synthesis the missing slice and recover the incorrect cardiac slice pose.

Quality-Aware Generative Adversarial Nets for Cross-Dataset Cardiac Cine MRI Synthesis.

Accurate ventricular volume measurements depend on complete heart coverage and correct cardiac orientation in CMR sequences that provide the most immediate indicators of normal/abnormal cardiac function. However, incomplete heart coverage, especially missing basal/or apical slices, and the slices in CMR sequences with incorrect cardiac orientation (ICO) are substantial problems that affect volume calculation, but are not sufficiently addressed in current clinical research. In this thesis, we propose two new deep architectures. One is called the missing slice imputation generative adversarial network (MSIGAN), which is used to learn the features of cardiac SAX slices across different positions and to consider the features as conditional variables to effectively infer missing

slices in query volumes. The other one is called unsupervised cycle-consistent adversarial network (SPSGAN), which provided with a SAX slice with ICO, automatically generates images under correct orientation. In a MSIGAN, slices are first mapped to latent vectors with position features through a regression net and then the latent vector with the desired position is projected onto the slice manifold, conditional on slice intensity, through a generator net. The latent vector preserved with the slice features (*i.e.*, intensity) and the desired position condition control generation versus. regression. Two adversarial networks are imposed on the regressor and generator, forcing the generation of more realistic slices. In a SPSGAN, we address this challenge by dividing the problem into two subtasks. First, we consider using a bidirectional generator that maps the initially rendered image back to an image with input cardiac orientation, which can be directly compared with the input image without requiring any GT images. Second, to generate high perceptual quality images, we propose a novel loss function that incorporates intensity and orientation terms.

In this thesis, we only focus on the cardiac ventricle coverage (point 1 in Figure. 1.5) and orientation (point 9 in Figure. 1.5). Although we have listed and explained the other criteria for CMR image quality evaluation in Section 1.2.3 and Figure. 1.5, there are also some other criteria that are easily to be assessed automatically, such as checking slice thick and gaps. Meanwhile, the artifacts (point 2 and 8 in Figure. 1.5) are less obvious in CMR image analysis.

Chapter 2

Deep Learning Methods in Cardiac Image Analysis

The limitation of hardware devices used in medical imaging systems impedes the diagnosis and treatment of heart-related pathologies. For example, the cardiac ventricular volume cannot be calculated directly using the imaging system, and various imaging equipment and the patient related reasons lead to poor image quality. To solve these problems and provide better clinical outcomes, computer-aided image analysis frameworks have been developed. Recently, significant progress has been made in DL. This progress is primarily due to the continuous improvement of computational capacity and the amount of available annotated data, as well as availability of improved DL models and algorithms. The essence of the application is to build a multi-hidden layer machine learning model, train the model using extremely large amounts of sample data, learn more accurate features, and ultimately improve classification or prediction accuracy. This chapter provides a brief overview of the previously developed analysis frameworks and their applications in medical imaging for classification, lesion detection, and segmentation across multi-modal images. These computation methods have the same goal, *i.e.*, time efficiency and objective quantitative analysis, as well as the evaluation, enhancement, and analysis of multi-modal imaging data.

In this chapter, we first focus on the principles of DL, highlight the popular CNNs and summarize image classification and segmentation frameworks. Then, we describe DL-based state-of-the-art medical image analysis methods. Finally, we discuss the challenges involved in practicable DL strategies for medical image analysis as well as open research directions.

2.1 Why Deep Learning?

DL has made significant progress in various fields and is the basis of artificial intelligence (AI) technology [123]. DL has helped achieve impressive results in medical image analysis [113]. In the current clinical workflow, medical image datasets are suitable for using DL technologies because of their large sizes compared with that of other medical data modalities, *i.e.*, text. DL can be applied to analyze various multimodal medical images, including X-ray, MRI, CT, and ultrasound as well as pathology and cell images. In addition, DL can help detect abnormal manifestations or lesions in a large number of data. For example, CT scan data have yielded positive results for larger lymph node and colonic polyp classification tasks [196]. In breast cancer screening, automatic classification of breast density using DL-based methods can help doctors predict the risk of breast cancer and prescribe the complementary screening [163]. In addition, DL technology can reduce the current high review rate for providing diagnostic support [3]. DL has helped achieve remarkable results in the research into new quantitative risk markers for breast cancer.

Traditional medical image analysis is primarily based on clinical experience, and traditional computation-based medical image analysis is primarily dependent on features that are manually extracted as features based on predefined calculation formulas. However, using manual features to describe medical images is extremely difficult because many meaningful image features are qualitative and extracted empirically. Using a data-driven method based on DL can can reduce the difficulty. For example, CNN models can automatically and autonomously extract and organize effective image features from large-scale labeled medical image data. Various CNN-based studies have provided evidence that low-level image features can be shared and fine-tuned among neural network models, such as transfer learning models. On the basis of this mechanism, medical image analysis can benefit from networks trained using a large number of natural images. For example, many studies have used ImageNet, which contains more than a million images to pretrain DL models, which are then effectively applied to medical image analysis tasks.

The emergence of DL will further the development of early screening techniques for diseases and deep mining of large population imaging databases will greatly facilitate the study of biomarkers. DL will also improve our ability to analyze and interpret large-scale datasets. In conclusion, recent developments in DL have shown a tremendous impact on medical image analysis and have helped achieve acceptable clinical levels using some important tasks that otherwise cannot be accomplished using non-DL methods. We expect that in the near future, research and clinical transformation in this area will flourish.

2.2 Conventional Deep Learning Models

This section discusses commonly used DL models, including stacked autoencoder (SAE), deep belief networks (DBNs), deep Boltzmann machines (DBMs), CNNs, and recurrent neural networks (RNNs). We focus on how various models learn multi-level image features from sample training data.

2.2.1 Supervised Learning Model: Convolutional Neural Networks

In 1989, LeCun proposed a CNN model in order to make better use of spatial structure information. CNNs can be used to capture visual local information because they take 2D or 3D image blocks as input. Typically, CNNs comprise several alternating convolutional and pooling layers, as well as fully-connected layers at the end of the network, as shown in Figure 2.1. Specifically, CNNs combine three different architectural concepts, *i.e.*, local receptive fields, weight replication (or shared weights), and spatial or temporal sub-sampling, to ensure the invariance of shift, scale and distortion. The feature maps that are approximately size-normalized and centered in a small neighborhood in the previous layer are received as the input plane to the current layer [124].

In a typical convolutional layer, the input is convoluted by convolution kernels and by adding bias terms. Finally, the feature map is generated using a nonlinear activation function. By denoting the i^{th} feature map of the l^{th} layer as $\boldsymbol{h}_i^l$ and the k^{th} feature map of the previous layer as $\boldsymbol{h}_k^{l-1}$, a convolution layer is formulated as:

$$\boldsymbol{h}_i^l = \sigma(\sum_k \boldsymbol{h}_k^{l-1} * \boldsymbol{W}_{ki}^l + \boldsymbol{b}_i^l), \tag{2.1}$$

where $\boldsymbol{W}_{ki}^l$ and $\boldsymbol{b}_i^l$ are the filter and bias terms that connect the feature maps between adjacent layers, $*$ denotes a convolutional operation, and $\sigma(\cdot)$ is an element-wise non-linear activation function [58].

Each feature map can be operated independently with pooling, which can gradually reduce the size of the representation space. Therefore, in CNN architectures, pooling layers are usually inserted between successive convolution layers to reduce the parameters and the computational burden in the network. Note that max pooling is the most common pooling method [214].

At the end of a convolutional network, typically, fully-connected layers are used for classification. In a CNN, all activations in the previous layer are connected to neurons in the fully-connected layer. Therefore, these activations can be calculated using matrix multiplication followed by a bias offset. A CNN is essentially an input-to-output mapping,

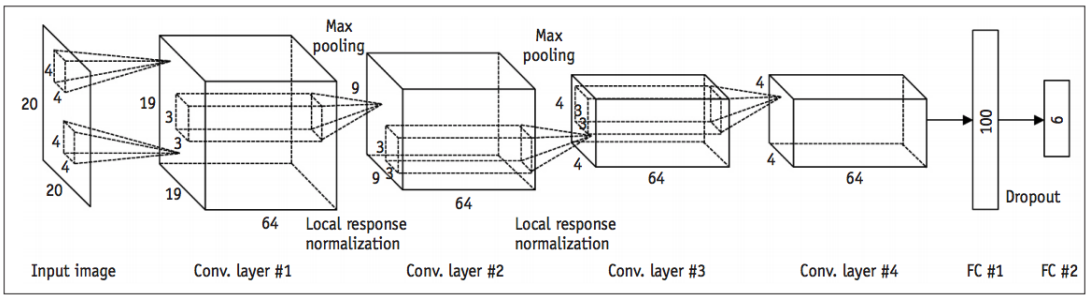


Figure 2.1: CNN architecture comprises convolutional, pooling and fully-connected layers. Each plane represents a feature map. This figure has been adapted from [124].

which enables the network to have input-to-output mapping capabilities via an end-to-end learning approach. Generally, training CNN network parameters is similar to training a traditional backpropagation algorithm. The output value is calculated by forward propagation. Then, the error between the output value and the ground-truth value is optimized using a gradient descent method to minimize the error, and finally the CNN parameters are adjusted by gradient backpropagation [252].

2.2.2 Unsupervised Learning Model: Stacked Autoencoder

An autoencoder is a type of unsupervised learning structure with an input layer, a hidden layer, and an output layer as shown in Figure 2.2. The process of training an autoencoder comprises an encoder and a decoder. The encoder is used to map the input data to a hidden representation, and the decoder is used to reconstruct the input data from the hidden representation. Given an unlabeled input dataset $\{\boldsymbol{x}_n\}_{n=1}^{N}$, where $\boldsymbol{x}_n \in R^{m\times 1}$, $\boldsymbol{h}_n$ represents the hidden encoder vector calculated from $\boldsymbol{x}_n$, and $\hat{\boldsymbol{x}}_n$ is the decoder vector of the output layer, the encoding process is expressed as follows:

$$\boldsymbol{h}_n = f(\boldsymbol{W}_1\boldsymbol{x}_n + \boldsymbol{b}_1), \tag{2.2}$$

where f is the encoding function, $\boldsymbol{W}_1$ is the weight matrix of the encoder, and $\boldsymbol{b}_1$ is the bias vector.

The decoder process is defined as follows:

$$\hat{\boldsymbol{x}}_n = g(\boldsymbol{W}_2\boldsymbol{h}_n + \boldsymbol{b}_2), \tag{2.3}$$

where g is the decoding function, $\boldsymbol{W}_2$ is the weight matrix of the decoder, and $\boldsymbol{b}_2$ is the bias vector. The autoencoders parameter sets are optimized to minimize the reconstruction

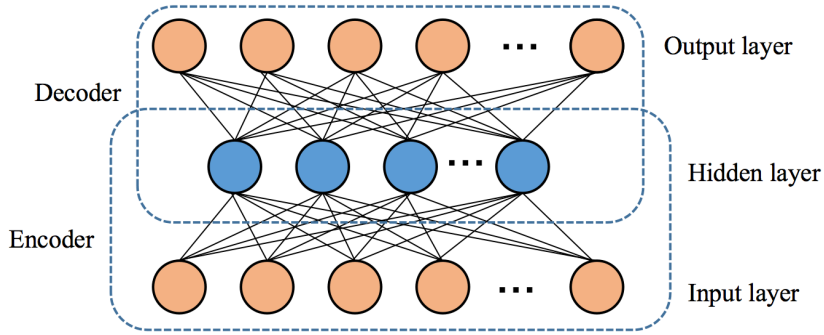


Figure 2.2: Structure of autoencoder.

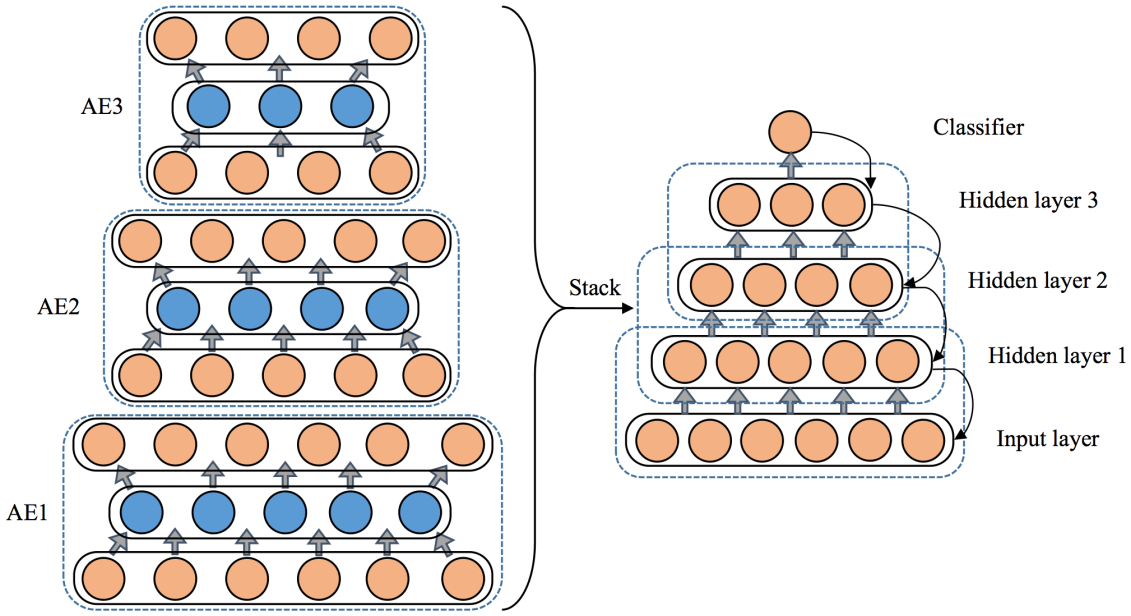


Figure 2.3: Structure of stacked autoencoders.

error:

$$\phi(\Theta) = \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{x}^i, \hat{\boldsymbol{x}}^i), \tag{2.4}$$

where L represents the loss function $L(\boldsymbol{x},\hat{\boldsymbol{x}}) = \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2$, $\boldsymbol{x}^i$ and $\hat{\boldsymbol{x}}^i$ are the input image and output image, respectively, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are the weight matrices of the encoder and decoder, respectively, n is the total of the training samples.

As shown in Figure 2.3, the SAE structure involves stacking n autoencoders into n hidden layers using an unsupervised layer-wise learning algorithm. The SAE structure is then fine-tuned using a supervised method [138]. The SAE-based method can be divided into three steps.

1. Train the first autoencoder using input data and obtain the learned feature vector;

2. The feature vector of the former layer is used as the input for the next layer. This procedure is repeated until training is finished.

3. After all hidden layers are trained, the backpropagation algorithm is used to minimize the cost function and fine-tune the weights using a labeled training set.

The SAE extracts the input image features from pixel-level data using an automatic coding-decoding network to improve the representation of the model. It has been widely used in dimensionality reduction and feature learning. The image data are not both the input and output of the SAE; thus it can detect whether the features learned by the middle layer of the network satisfy the requirements. If we restrict sparsity to each layer in the SAE, we can obtain a stacked sparsely autoencoder (SSAE), which can provide the model with a certain anti-noise ability and better generalization [216] [247]. When input images are represented by the SSAE, different network layers represent different levels of features, *i.e.*, the lower layers of the network represent simple patterns, and the higher layers represent intrinsic patterns that are more complex and abstract in the input vectors.

2.2.3 Fine-tuning Network

For a given a DL task, *e.g.*, a classification problem that involves training a CNN (ConvNet) model on the ImageNet dataset, our first instinct is to begin training the network from scratch. However, deep neural networks (*e.g.*, ConvNet) have a large number of parameters, typically in the range of millions. It is difficult to train deep-seated neural networks from scratch (*i.e.*, complete training). First, training ConvNet from scratch is difficult because it requires a large amount of labeled training data and expertise to ensure proper convergence [231]. Second, training deep CNN requires significant computing and memory resources, and without sufficient resources, the training process will be very time consuming [60]. Training Covnet on small datasets (smaller than the number of parameters) significantly affects its generalizability and frequently leads to overfitting. Therefore, it is tedious and time consuming to train the network from scratch, which requires diligence, patience and professional knowledge.

Commonly, researchers fine-tune existing networks via continual training using backpropagation. Pretrained networks are typically trained on a large-scale dataset (e.g., ImageNet; 1.2 million labeled images) and have been successfully applied in many computer vision tasks, such as feature extractor or as a baseline in transfer learning [212] [11] [180]. If there is no significant difference between the natural image dataset and the context of

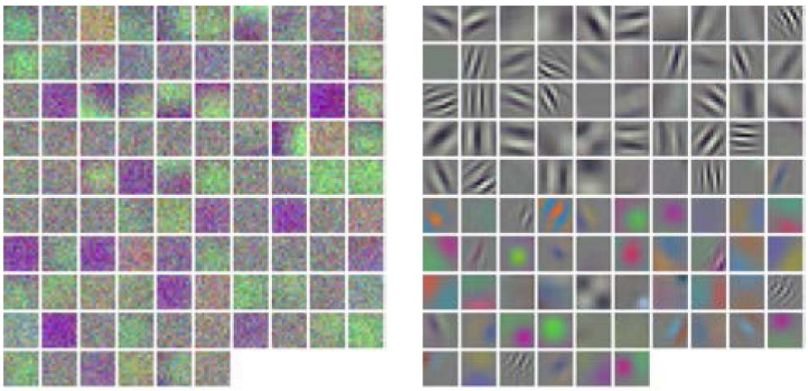


Figure 2.4: Filter of AlexNet's first layer by (a) training from scratch on interstitial lung diseases (ILD) CT scans data, (b) fine-tuning pretrained on ImageNet version. This figure has been adapted from [5].

the original dataset (*e.g.*, ImageNet), the pretrained model demonstrates learning characteristics related to the target classification problem. Generally, if a given dataset does not differ from the dataset used by the pretrained model, we should fine-tune the network. If the target datasets come from very specific domains, such as medical images or handwritten Chinese words, we should consider training a deep network from scratch. There are several additional considerations. 1) If the target dataset is small, fine-tuning the pretrained network on this small dataset will lead to over fitting, especially in case of the VGG network, which only has a few fully-connected layers in the last few layers of the network. 2) If the target dataset has several thousand samples, fine-tuning could achieve a better result using common data augmentation methods, such as translation, rotation, and flipping. 3) If the target dataset is very small (less than 1,000 samples), we could preselect the output of intermediate layers prior as bottleneck features and input them to a linear classifier (*e.g.*, an SVM). An SVM is particularly effective in identifying decision boundaries on a small dataset.

In the medical image processing domain, recent transfer learning methods can be divided into two categories. The first category is typically considered a pretrained CNN as a feature extractor [13] [243] [6]. For example, we can take an image as the input to a pretrained CNN and extract the features from certain layers of the network as the input of a new pattern classifier. This method has been applied widely. For example, Bar et al [13] used pretrained CNNs as a feature extractor for chest pathology identification. Another study [243] demonstrated that combining CNN-based and handcrafted features can improve the performance of dedicated nodule detection even though the pretrained CNNs demonstrate worse performance on it.

The second category involves adapting pretrained CNNs to target application. For example, Carneiro et al. [22] replaced the fully-connected layers in pretrained CNNs with a new logistic layer. Then, they retained the rest of the network and trained the new network using labeled data. This method helped achieve promising results in the classification of unregistered multi-view mammograms. In addition, a fine-tuned pretrained CNN has been utilized in the localization of the standard planes in ultrasound images [30]. Gao et al. [71] fine-tuned all layers of a pretrained CNN using an attenuation re-scale scheme for automatic classification of interstitial lung diseases. The attenuation re-scale scheme converts single-channel CT slices to RGB-like images, which is required to fine-tune a pretrained model. Ciompi et al. [39] proposed the automatic detection of pulmonary fissure nodules using a pretrained CNN by ImageNET and fine-tuning the network using a small number of labeled CT data sequences. Tajbakhsh [231] demonstrated that the performance of deep fine-tuning is better than that of shallow fine-tuning, and that the importance of fine-tuning a network is enhanced when the size of the training dataset is reduced. Differing from the above approaches, Schlegl et al. [207] utilized a fine-tuning method in an unsupervised network. They developed unsupervised approaches to pretrain CNNs and injected information from images and sites without annotations. This type of cross-site pretraining demonstrated improved classification compared with the methods that initialize model parameters randomly.

Some general guidelines for fine-tuning implementation are summarized as follows.

1. The common approach is to drop out the last layer (*e.g.*, the softmax layer) of the pretrained network. This dropped layer is then replaced with a new logistic layer that is related to the target problem. For example, a softmax layer with 1000 classes is connected to the pretrained network with the ImageNet dataset. If the target problem is a classification task with 10 categories, the original 1000 categories of the softmax layer are replaced with softmax layers with only 10 categories. Then, the backpropagation method is used to fine-tune the pretrained network's parameters. Meanwhile, we should employ cross-validation to ensure that the new network can generalize well.

2. Using a smaller learning rate, such as 10 times smaller than that used for scratch training, to train the network is beneficial. Here we expect the parameters of the pretrained network to be sufficient compared with randomly initialized parameters and we do not want to distort them too quickly or too much.

3. Another common approach is to freeze the parameters in the first few layers of the pretrained network. On one hand, we wish to keep the parameters constant because

the first few layers can capture general features, such as the curves and edges that are relevant to the target problem. On the other hand, we will make the parameters in the subsequent layers more related to the dataset-specific features.

2.3 Generative Adversarial Networks

Generative adversarial network (GAN) [81] has become a hot research topic in the field of AI, especially the DL. The basic idea of a GAN comes from two-person zero-sum game theory, which compromises a generator and a discriminator and is trained using an adversarial approach. The aim is to estimate the potential distribution of data samples and to generate new data samples from the same distribution. In the fields of image and visual computing, voice and language processing, information security, chess games, etc., GANs have been widely studied and have great application prospects. In this section, we focus on the research progress and prospects of GANs, and summarize the background, theory and implementation model; application fields; advantages and disadvantages; and development trends of GANs.

2.3.1 The Conventional GAN Model

The core idea of a GAN [81] comes from game theory. It sets the players as a generator and a discriminator. The purpose of the generator is to learn the distribution of real data as much as possible, whereas the purpose of the discriminator is to distinguish as correctly as possible whether the input data is from the real data or from the generator. To win the game, the two players need to constantly optimize and improve their ability to generate and discriminate. Figure. 2.5 shows the structure of a GAN. We use differentiable functions D and G to represent the discriminator and the generator, respectively. Their inputs are real data $\boldsymbol{x}$ and random variables $\boldsymbol{z}$, respectively. $G(\boldsymbol{z})$ is the samples generated by G that match the distribution of real data as much as possible. If the input of the discriminator comes from real data, the output of D is 1, otherwise the output is 0. Here the goal of D is to achieve the binary classification for data sources: true (from the distribution of real data $\boldsymbol{x}$) or false (from the generator's fake data $G(\boldsymbol{z})$). The goal of G is to make the distribution $D(G(\boldsymbol{z}))$ of the generated fake data $G(\boldsymbol{z})$ on D match with that of real data $\boldsymbol{x}$ on $D(\boldsymbol{x})$. These two processes of confrontation and iterative optimization improve the performance of D and G. When the discriminant ability of D improves to a certain extent and the data source cannot be correctly identified, it means that G has learned the distribution of real data [185].

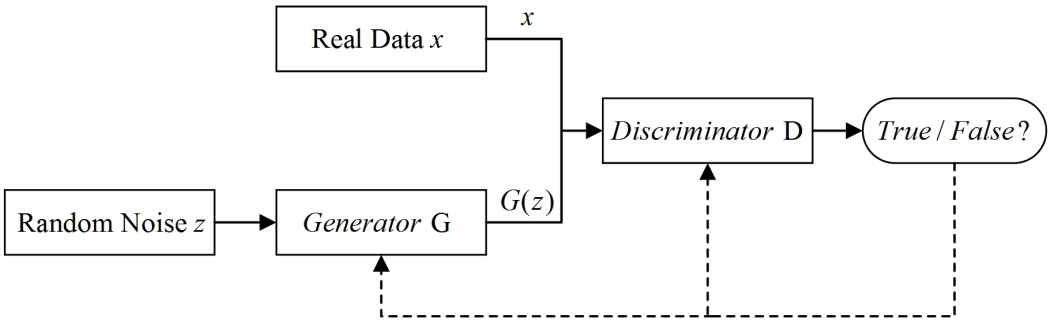


Figure 2.5: Computation procedure and structure of GAN

First, given the generator G, we consider that the optimization of the discriminator D is also a process of minimizing cross-entropy based on sigmoid. The loss function is defined as follows:

$$Obj^D(\theta_D, \theta_G) = -\frac{1}{2}E_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] - \frac{1}{2}E_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(g(\mathbf{z})))] \tag{2.5}$$

where x is sampled from the real data distribution $p_{data}(x)$, z is sampled from the prior distribution $p_z(z)$, *e.g.* the Gaussian noise distribution. The training dataset of the discriminator comes from two parts: the real dataset distribution $p_{data}(x)$ (labeled as 1) and the generated data distribution $p_g(x)$ (labeled as 0). θ_D and θ_G are the learned parameters in D and G, respectively. Given a generator G, we need to minimize Eq. (2.5) to obtain the optimal solution.

On the other hand, $D(x)$ represents the probability that x comes from real data or generated data. When the input data is sampled from the real data x, the goal of D is to make the output probability value $D(x)$ approach 1. When the input of D comes from the generated data $G(z)$, the goal of D is to correctly classify the data, so that $D(G(z))$ approaches 0, and the goal of G is to make the generated data approach 1 [8]. This is actually a zero-sum game about G and D, so the loss function of G is $Obj^G(\theta_G) = -Obj^D(\theta_D, \theta_G)$. So the optimization problem of a GAN is a min max problem. The objective function of a GAN can be described as follows:

$$\min_G \max_D \{f(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]\} \tag{2.6}$$

In summary, during the GAN learning process, we need to train D to maximize the accuracy of its discriminating ability that distinguishes whether the input data comes from real data or the generated data distribution $G(z)$. On the other hand, we need to train G to minimize log(1-D(G (z)). An alternate optimization method can be used for the entire

process: fix G and optimize D to maximize the discriminating ability of D and then fix D and optimize G to minimize the discriminating ability of D. When $p_{data} = p_g$, the model is optimized using the global optimal solution.

2.3.2 GAN based derivative models

Since Goodfellow et al. [81] proposed a GAN in 2014, various GAN based derivative models have been proposed. The innovations in these models include model structure improvement, theoretical expansion, and application.

Odena et al. [173] proposed semi-GAN, which added the annotation information of real data to the training process of the discriminator D. Furthermore, conditional GAN (CGAN) [162] was proposed by adding additional information y to G, D and real data to model, where y can be labels or other auxiliary information. Traditional GANs learn a generative model to map the data distribution of hidden layers to the distribution of complex real data. Donahue et al. [55] proposed a bidirectional GANs (BiGANs) to map complex data to the space of hidden layers so as to realize feature learning. In addition to the basic framework of GANs, BiGANs add an additional decoder Q to map real data $\boldsymbol{x}$ to hidden layer space, and its optimization problem is converted to $\underset{G,Q}{min}\underset{D}{max} f(D,Q,G)$.

Information GAN (infoGAN) [34] is another important extension of a GAN. A GAN can learn effective semantic features, but the relationship between the input noise variable $\boldsymbol{z}$ and the specific semantic meaning is not clear. An infoGAN can obtain mutual information between input variables and specific semantics. The specific implementation is to divide the input of generator G into two parts: $\boldsymbol{z}$ and $\boldsymbol{c}$, where $\boldsymbol{z}$ is the same as the input of the GAN and c is used to represent the implicit relationship between structural hidden variables and specific semantics. A GAN sets $p_G(\boldsymbol{x}) = p_G(\boldsymbol{x} \mid \boldsymbol{c})$, but in fact, $\boldsymbol{c}$ and the output of G have a strong correlation. $G(\boldsymbol{z},\boldsymbol{c})$ is used to represent the generator's output. The authors in [34] proposed that mutual information $I(\boldsymbol{c};G(\boldsymbol{z},\boldsymbol{c}))$ is used to represent the correlation of $\boldsymbol{c}$ and G, and the objective function is

$$\underset{G}{min}\underset{D}{max} f_I(D,G) = f(D,G) - \lambda I(\boldsymbol{c};G(\boldsymbol{z},\boldsymbol{c})) \tag{2.7}$$

Odena et al. [174] proposed the auxiliary classifier GAN (AC-GAN), which can realize multi-classification problems. The discriminator outputs the corresponding label probability. In practical training, the objective function contains the likelihood of the real data source and of the correct classification label. It can adjust the loss function further so that the classification accuracy is higher. The key of an AC-GAN is that the corresponding image label can be generated using the label information of the generator, and simultaneously

it can expand and adjust the loss function, which can further improve the generating and discriminating ability of the GAN.

Considering that a GAN's output is a continuous real number distribution and cannot generate a discrete space distribution, Yu et al. [268] proposed a generative model SeqGAN, which can generate a discrete sequence. They used an RNN to implement generator G, a CNN to implement discriminator D, and the probability output of D to update G through reinforcement learning.

2.3.3 Advantages and Disadvantages of GANs

GANs are of great significance to the development of generative models. As a generative method, a GAN effectively solves the problem of data generation that can be interpreted naturally. Especially for high-dimensional data generation, the adopted neural network structure does not restrict the data generation dimension, which greatly broadens the scope of generating data samples. The neural network structure can integrate all kinds of loss functions and increase design flexibility [204]. The adversarial training method abandons the direct replication or mean of real data and increases the diversity of generated samples. It is easy to understand these generated samples in practice. For example, the ability to generate sharp, clear images provides a possible solution to creatively generate data that is meaningful to humans.

GANs not only contribute to generative models but also inform semisupervised learning. The GAN learning process does not require data labels. Although a GAN is not designed for semisupervised learning, the training process of a GAN can be used to implement the pretraining process with unlabeled data in semisupervised learning [185]. In particular, a GAN is pretrained with unlabeled data to understand the data. Then a small number of labeled data items are used to train the discriminator for traditional classification and regression tasks.

GANs solve some problems of generative models and help in the development of other methods, but GANs are not perfect. They also introduce some new problems when solving the existing ones. A GAN adopts the criterion of AL, and it is theoretically difficult to judge the convergence of the model and existence of equilibrium points. The training process should ensure the balance and synchronization of the two adversarial networks, otherwise it is difficult to achieve good training results [174]. In practice, the synchronization of two adversarial networks is not easy to control, and the training process may be unstable. In addition, a GAN is a kind of generative model that is based on a neural network; it has the general disadvantage of a neural network model, i.e., its interpretability is poor.

Although GANs exhibit these problems, it is undeniable that the progress in the research on GANs has shown that GANs have broad prospects for development. For example, the Wasserstein GAN [8] completely solved the problem of training instability and solved the phenomenon of collapse mode [80]. How to thoroughly solve the collapse mode phenomenon and optimize the process are the research directions of GANs. In addition, the theoretical inference of GAN convergence and the existence of equilibrium points are important research topics for the future. These research directions aim to better solve the disadvantages of GANs. From the perspective of development and application of GANs, how to generate diverse and interactive data from a simple random input is a recent direction for application development. From the perspective of the cross-integration of GANs and other methods, how to better integrate a GAN with feature learning, imitation learning, reinforcement learning and other technologies, develop new AI applications; and promote the development of these methods are meaningful directions. In the long run, how to use GANs to promote the development and application of artificial intelligence (AI), to enhance the ability of AI to understand the world, and to even stimulate the creativity of AI are questions worthy of consideration.

2.4 Deep Learning in Medical Image Analysis

In recent years, DL has demonstrated excellent performance for natural image processing in computer vision and has facilitated breakthroughs in medical image analysis. Currently, most scholars in this field explore CNNs for image classification, which is primarily used for medical image analysis tasks, such as lesion recognition, detection and segmentation. In 2016, an IEEE Trans on Medical Imaging special issue on DL presented recently developed CNN architectures and DL applications in medical imaging processing. This special issue contains 18 papers by various research scholars from around the world. A variety of classical tasks were presented, including detection and classification problems, such as lesion detection, image segmentation, shape modeling, and image registration. In addition, some new application domains were proposed in these studies. With the exception of classical tasks, the exploration of networks and insight into these architectures were also included for some specific tasks, parameters, the selected training sets, and more [82]. Table 2.1 lists the classical CNN-based frameworks for computer vision classification tasks. In this section, we first introduce the CNN framework for classification and segmentation in medical image analysis. Then, we summarize the research status of DL in medical image classification, detection, and segmentation and other applications.

Table 2.1: Classical CNN frameworks for computer vision classification tasks

Model	**Novelty**	**Application, Remarks**
LeNet [124]	Multiple convolution layers and sub-sampling layers	US handwritten digit recognition
AlexNet [118]	Proposed ReLU and Dropout	Refreshed the world record of the 2012 ImageNet ILSVRC object classification competition
VGGNet [218]	Using small convolution kernels achieve deeper network and multi-scale fusion	The ILSVRC 2014 localization task champion, classification task runner-up
GoogleNet [229]	Network with 22 layers and one more Inception in series	The ILSVRC 2014 detection and classification tasks champion
ResNet [85]	Introducted the residual network, and the skip connection, 152 layers	Object detection and object recognition champion in the 2015 ILSVRC competition
Inception ResNet [228]	Combined the architecture of Inception and Residual Net	The performance is comparable to ResNet and the speed of convergence is faster
FCN [144]	Pixel level classification is achieved in dense prediction	The problem of repeated convolution calculation is avoided from overlapping of image blocks
DenseNet [96]	There is a direct connection between any two layers	Alleviating gradient disappearance, enhancing feature propagation, supporting feature reuse, and reducing the number of network parameters
SqueezeNet [98]	Simplifying network structure and reducing network parameters	Only need 1/50 parameters in AlexNet can achieve the same accuracy as AlexNet
DCNN [45]	A deformable deep convolution neural network is proposed	Enhance the modeling ability of network for geometric transformation
DPN [35]	Combined the advantages of ResNet and DenseNet	Object detection and object recognition champion in the 2017 ILSVRC competition
SENet [94]	Learn the importance of each feature channel and enhance useful features	2017 ILSVRC image classification competition champion

2.4.1 Medical Image Classification

(1) Image screening

Image screening is one of the earliest applications of DL in medical image analysis. It involves taking one or more examination images as the input, predicting them using trained models, and outputting a diagnostic variable that indicates whether a disease or severity has been classified [136]. Image screening is a type of image-level classification, and the DL models used to solve this task initially focus on SAE, DBN, and DBM networks and unsupervised pretrained methods. Many studies have examined neuroimaging analyses, such as the diagnosis of Alzheimer's disease (AD) or mild cognitive impairment [225] [226] [141]. These algorithms typically use multimodal images as the input to extract complementary feature information from MRI, PET, and CSF.

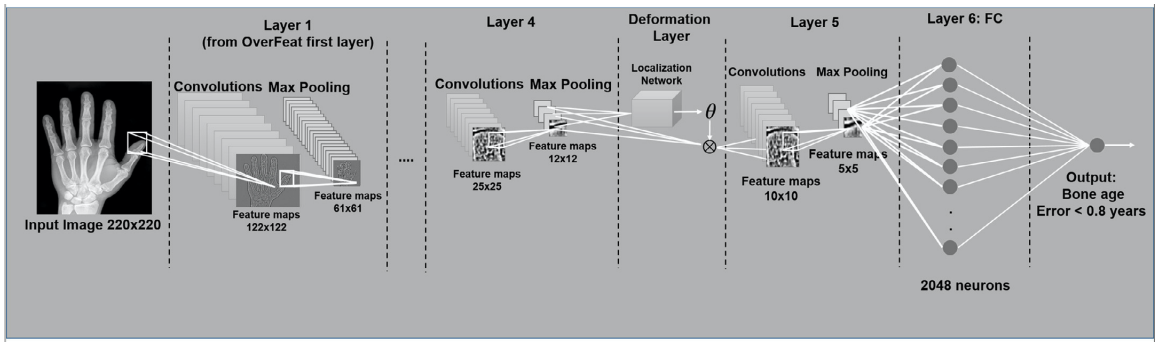


Figure 2.6: Overview of BoNet architecture. The architecture comprises five convolutional and pooling layers (to extract low and middle-level visual features) one deformation layer facing bone nonrigid deformation and two fully connected layers for bone age regression. This figure has been adapted from [222].

Recently, CNNs have been applied in many fields, and they have gradually become a standard image classification technology. For example, Arevalo et al. [7] proposed a representation learning framework for breast cancer diagnosis. This framework automatically uses a CNN to learn discriminate features to classify breast X-ray lesions. Kooi et al. [116] compared traditional handcrafted feature extraction and automatic CNN feature extraction methods, where both methods were trained on a large dataset (45,000 mammograms). The results indicated that the CNN extracted features were superior to the traditional handcrafted features at low sensitivity and that the two methods were equivalent at high sensitivity. Spampinato et al. [222] applied a deep CNN to automatically evaluate skeletal age (as shown in Figure 2.6). Xu et al. [263] studied the classification of colon cancer from histopathological images using a deep CNN and multi-instance learning method to extract features automatically with only a few manual annotations. Gao et al. [74] discussed the

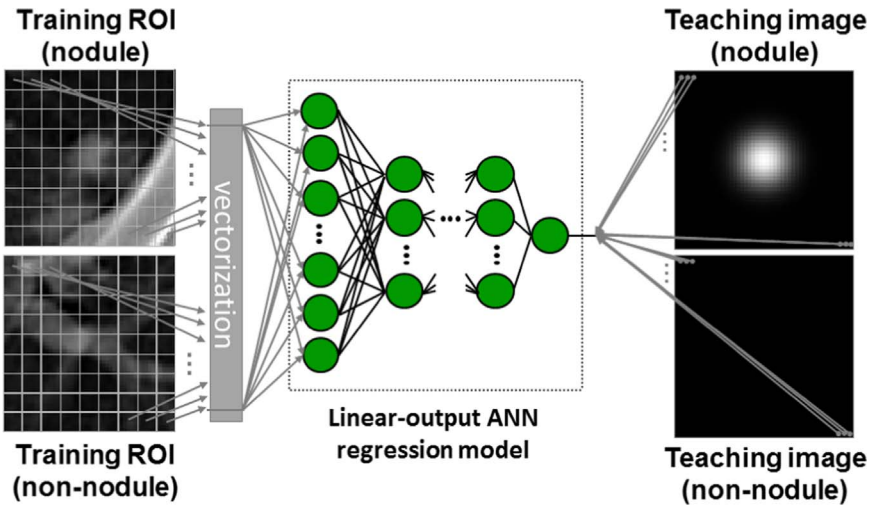


Figure 2.7: Schematic overview of massive-training artificial neural networks (MTANN) training. Non-overlapping patches are depicted in the region of interest (ROI) to avoid clutter. Image patches are extracted densely from each ROI, which results in a massive set of training patches. This figure has been adapted from [232].

importance of DL techniques, particularly CNNs, in brain CT image classification tasks, which provide supplementary information for early diagnosis of AD. Payan et al. [179] and Hosseiniasl et al. [93] used a 3D CNN to diagnose AD in neuroimaging. In addition, Abdi et al. [2] used a CNN to automatically assess echocardiogram quality (apical four-chamber view). Gao et al. [72] combined two 2D CNNs to extract the temporal and spatial features of echocardiograms, and then classified the viewpoints of echocardiograms to help diagnose heart disease.

In addition, some studies have combined CNNs and RNNs for medical image screening tasks. For example, Gao et al. [73] adopted a CNN to extract low-level local feature information from slit lamp images and further extracted high-level features using an RNN to classify nuclear cataracts.

(2) Object or lesion classification

Object or lesion detection and classification can assist doctors in disease diagnosis, such as the classification of benign or malignant breast lesions. This process first identifies or marks the specific region using image preprocessing methods. Then, objects or lesions in the specific region are classified. Accurate classification requires both local information about lesion appearance and global context information about the location.

CNN-based frameworks are widely used in lesion classification tasks. For example, Anthimopoulos et al. [5] used a CNN to design a multi-classification framework to distinguish the patterns of interstitial pulmonary diseases, such as ground-glass disease, honey-

comb disease, calcification, and pulmonary nodules. This method helped achieve 85.5% accuracy. Kawahara et al. [112] used a multi-processing flow-based CNN to classify skin lesions, where each flow processed images at different resolutions. Jiao et al. [104] used a CNN to extract deep features at different levels to improve classification accuracy for breast cancer. Tajbakhsh et al. [232] detected lung nodules on CT images to distinguish benign and malignant pulmonary nodules. In addition, they compared the performance of massive-training artificial neural networks (MTANNs) and CNNs (as shown in Figure 2.7). The experimental results demonstrated that the MTANNs performance was much better than that of the CNNs when using limited training data.

Some researchers have combined a CNN with other basic models to achieve good performance in medical image classification tasks. For example, Kallenberg et al. [107] combined a CNN and an SAE to construct the convolutional sparse autocoder (CSAE) model. Then, they used an unsupervised pretrained CSAE model to perform breast density segmentation and breast risk assessment. Van et al. [245] combined the discrimination ability of a CNN and the generation ability of the restricted Boltzmann machine (RBM) to construct a CRBM to analyze lung CT images. Zhang et al. [275] developed a two-layer CNN architecture comprising a pointwise gated Boltzmann machine and an RBM for shear-wave elastography feature extraction. Compared with statistical features that quantify image intensity and texture, deep features learned by a CNN helped achieve a better classification performance (93.4% accuracy). Shi et al. [215] used a new deep polynomial network to classify a small number of ultrasound datasets, and the classification accuracy of the chest and prostate datasets was 92.4% and 90.28%, respectively, which are better than the results achieved using DBN and SAE-based methods.

2.4.2 Object or Lesion Localization and Detection

Accurately locating specific biomarkers or anatomical structures in medical images is of great significance in clinical treatment and is directly related to treatment effects. To process 3D data using DL algorithms, in some methods, the 3D space as a combination of 2D orthogonal planes such that the location task can be converted into a classification task and processed using a general DL framework. For example, Yang et al. [265] combined CNN-learned features from three orthogonal directions to identify the markers of the femoral end, where the 3D position of the marker is defined as the intersection of three 2D images. Chen et al. [29] extended the fully-convolutional network (FCN) to a 3D FCN and proposed a 3D FCN-based localization and segmentation method that helped achieve very good results in the 2015 vertebral disk localization and segmentation challenge. Vos et al. [50] identified the 3D rectangular bounding box by resolving the 3D CT volume into a 2D

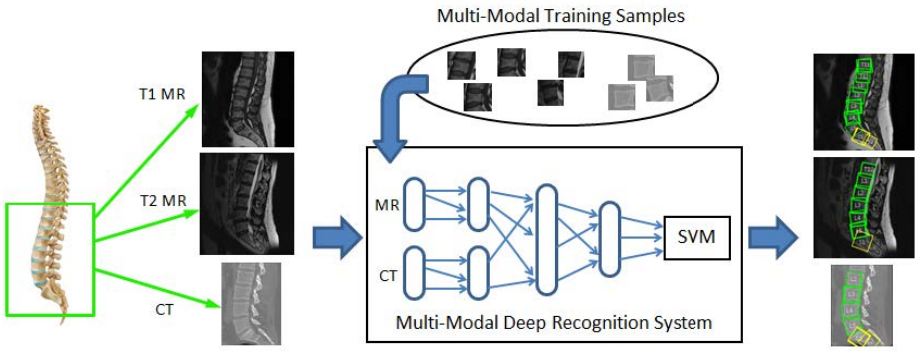


Figure 2.8: Multi-modal recognition for lumbar spine imaging. The modalities are uniformly trained and detected in one unifed recognition system, in which features from different modalities are fused and enhanced by each other via a deep network. This figure has been adapted from [20].

form and locating the anatomical regions of the heart, aortic arc and descending aorta of interest. In addition, LSTM was used to process the time information contained in medical videos. Kong et al. [115] combined an LSTM-RNN and CNN to detect ED and ES frames in cardiac MRI videos. Cai et al. [20] [127] used a deep CRBM to extract and fuse image features from different modalities in an unsupervised manner in order to identify vertebrae in MR and CT images.

Detecting the ROI or lesion in medical images is important for diagnosis. It has a long research history in the development of computer-aided detection systems. This work is typically designed to detect lesions automatically to improve detection accuracy or reduce expert reading time [136]. The implementation process comprises two steps, *i.e.*, locating the ROI in the entire image space and identifying small lesions in the ROI.

As early as 1995, Lo et al. [142] proposed the first object detection system using an ANN that utilized a four-layer CNN to detect nodules in X-ray images. Ciresan et al. [42] successfully detected mitotic cells in breast cancer pathological tissue images using a deep CNN as a pixel classifier. Sirinukunwattana et al. [219] used a spatially constrained CNN and neighborhood ensemble predictor to improve the accuracy of detecting and classifying colon cancer cell nuclei on pathological images. Li et al. [128] proposed a glaucoma detection method based on a deep CNN classification network. Roth et al. used a deep CNN to improve the object detection accuracy on CT images. Their main idea was to extract ROI candidates using existing methods, and then learn the high-level features of the object based on the deep CNN in order to ultimately detect and segment the object using the learned features. They significantly improved the object detection accuracy of several applications, such as automatic lymph node detection on abdominal CT images

[197] and sclerotic metastasis and colon polyp detection [196] [198]. Wang et al. [249] used a 12-layer CNN to detect breast artery calcification in mammograms. The results of a quantitative analysis of calcium quality demonstrated that the detected calcium quality was close to the gold standard, and the accuracy reached 96.24%. Quellec et al. [190] proposed a solution to automatically detect referable diabetic retinopathy (DR) and DR-related lesions. Referable DR is detected at the image-level using trained ConvNets; furthermore, the pixels that play a role at the image-level predictions are detected. Finally, a heatmap with the size of the image is visualized and obtained. To produce high quality heatmaps, *e.g.*, reducing attenuation artifacts, they proposed enhancing the sparsity of the heatmaps while training the ConvNets; good results were achieved using this method in the 2015 Kaggle DR competition.

Pixel classification is the key to detect image ROIs or lesions. Currently, most DL-based object detection systems use a CNN to perform pixel classification tasks, followed by using postprocessing methods to obtain the object. CNN-based frameworks and methods are similar to general pixel-level classification methods and must be combined with the neighborhood context or 3D information of the image to improve classification accuracy, such as integrating multi-view information [221] or multi-modal images [234] using multiple CNNs. Albarqouni et al. [4] used a multi-scale CNN method to detect mitosis on breast cancer pathological images. Chen et al. [32] approximately expressed the features of 3D medical images using 2D deep features combined with an SVM classifier, which realized the automatic detection of cerebral microbleeds (CMB) using susceptibility weighted imaging. Dou et al. [58] improved the work reported in the literature [32] by adopting a cascaded 3D CNN framework to make full use of the spatial context information in MR images to extract high-level features that can better represent CMBs. This method was verified extensively on a large dataset with 320 MR images and achieved high sensitivity (93.16%). This group also used a multi-level 3D CNN framework to detect pulmonary nodules on CT images, which was validated by the Luna 16 Challenge in ISBI 2016. The algorithm achieved the highest results relative to reducing false positive indicators [57]. Van Grinsven et al. [244] used a CNN to extract features and adopted a positive and negative sample equalization strategy to effectively detect hemorrhage on color fundus images.

Several studies have used other DL methods to achieve interesting object or lesion detection results. For example, Shin et al. [216] applied an SAE to detect abdominal organs on MRI. First, spatial features were learned in an unsupervised manner; then, multi-organ detection was performed based on the learned 'interest points'. Xu et al. [261] used an SSAE network to learn deep features from histopathological images for breast cancer nuclei identification to determine the stages of breast cancer. Masood et al. [157] proposed

a semi-supervised learning algorithm based on a DBN and SVM to recognize dermoscopic melanoma automatically. This solution can help solve problems when a limited amount of labeled training data is available. Differing from the traditional CNN method, Li et al. [131] used a Sobel edge contour and a Gabor texture feature as input and adopted a CNN for feature fusion and deep feature extraction, which improved automatic detection accuracy for lumbar vertebrae from C-arm X-ray images.

Recently, some studies have applied CNN-based methods to develop detection and localization tools in surgical videos. For example, Girshick et al. [79] and Sarikaya et al. [205] proposed architectures using multi-modal CNNs for fast detection and localization tools for understanding robot-assisted surgery videos. Twinanda et al. [239] designed a new CNN-based framework (EndoNet) to learn visual features automatically from cholecystectomy videos, while performing phase recognition and detection tasks in a multitask manner. Chen et al. [31] proposed a method that combines a CNN and LSTM to detect multiple standard planes in ultrasound images automatically, which is helpful for substantive biometry and diagnosis.

2.4.3 Medical Image Segmentation

(1) Organ and tissue segmentation

Segmentation of organs and their substructures in medical images can be used to quantitatively analyze clinical parameters that are related to their volume and shape (*i.e.* ventricular volume and EF of the heart). DL is widely used in these tasks, such as cardiac ventricle segmentation, vascular segmentation, histopathology and microscopic image segmentation.

LV segmentation from cardiac MRI is an important step in calculating the ventricular volume and EF of the heart. Manual contouring is a tedious, operator-dependent, and time-consuming task; thus, researchers have studied semiautomatic and automatic ventricle segmentation approaches to obtain consistent and accurate delineations from SAX and LAX images, as shown in Figure 2.9. Carneiro et al. [23] used a DBN to learn the features and model the appearance of the LV. They then segmented the LV automatically on an ultrasonic image of the heart using a supervised learning approach. Avendi et al. [10] used SAE learned deep features to preliminarily infer the shape of the LV and then combined with a deformation model to improve the accuracy and robustness of LV segmentation. Ngo et al. [169] combined a DBN using the level-set method to segment the LV from cardiac MRI automatically. The FCN-based deeply supervised network framework and conditional random field refinement method proposed by Dou et al. [59] have achieved state-of-the-art

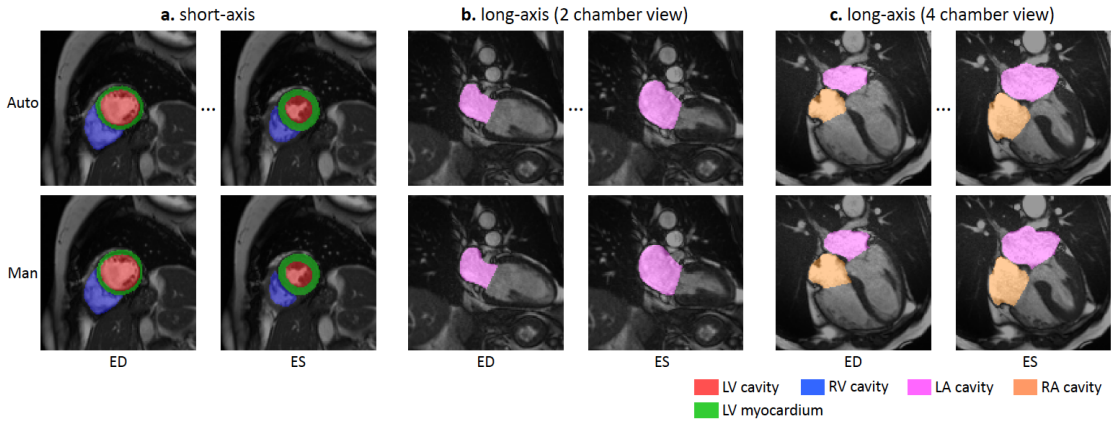


Figure 2.9: Examples of automatic cardiac MR image (SAX and LAX images) segmentation results obtained using a CNN. The top row shows the automated segmentation results for all ED and ES frames. The bottom row shows the manual segmentation. Manual analysis only annotates ED and ES frames; thus, the automated method only shows the ED and ES frames. The cardiac chambers are represented by different colors. The number of pixels labeled as BP and myocardium classes is calculated to obtain clinical measurements, such as EF and ventricular mass. This figure has been adapated from Ref. [12].

performance in heart and aorta segmentation. Tan et al. [233] parameterized the complete (all short axis slices and phases) LV segmentation task in terms of the radial distances between the LV center point and the endo- and epicardial contours in polar space. They then utilized CNN regression to infer these parameters. Zhen et al. [195] used a multi-scale CRBM for unsupervised learning, and then trained regression forest predictors using labeled data to estimate the biventricular volume directly from MR images.

DL has also been applied to challenging vascular segmentation tasks. For example, Nasr-Esfahani et al. [168] proposed the use of a CNN to detect vessel regions in angiography images, and Wu [257] presented a generic approach for vascular structure identification from medical images, which can be used for multiple purposes. This proposed method uses the state-of-the-art deep CNN to learn the appearance features of the target. A principal component analysis-based nearest neighbor search is then utilized to estimate the local structure distribution, which is further incorporated into the generalized probabilistic tracking framework to extract the entire connected tree. Liskowski [135] proposed a supervised segmentation technique that uses a deep neural network trained on a large (up to 400,000 images) dataset. The networks significantly outperform previous algorithms with respective to the area under ROC curve metric (up to > 0.99) and classification accuracy (up to > 0.97). Wang [251] proposed a supervised method that combined a CNN and the random forest to solve the segmentation problem relative to retinal vascular disease. Most of these methods are supervised feature extraction approaches that are combined with other

existing techniques and classifiers to improve segmentation accuracy. Differing from such classification-based segmentation methods, Li et al. [129] remolded the segmentation task as a problem of cross-modality data transformation from a retinal image to a vessel map, where a wide and deep neural network with strong induction ability was used to model the transformation. The experiment results demonstrated that it is an efficient training strategy.

Computer-assisted image feature extraction from surgical and biopsy specimens can benefit the prediction of the extent of disease aggressiveness and can be further used for disease diagnosis and classification. The key components of such predictors are image features extracted from histopathological images [153]. Currently, most segmentation methods for histopathological and microscopic images are based on DL, and many studies have achieved excellent segmentation results using CNN-based networks. Ciresan et al. [40] used a special type of deep artificial neural network as a pixel classifier for automatic segmentation of neuronal structures in stacks of electron microscopy images. The label of each pixel (membrane or nonmembrane) is predicted from raw pixel values in a square window centered on it. Kumar et al. [121] proposed a DL-based technique in which a CNN is used to produce a ternary map for segmenting nuclei from hematoxylin and eosin (H&E) pathological images. Xu et al. [262] transferred features extracted from CNNs trained using a very large general image database (ImageNet) to the medical image challenge and achieved 97.5% classification accuracy and 84% segmentation accuracy in the MICCAI 2014 Digital Pathological Challenge of Brain Tumors. Qaiser et al. [189] used a CNN to extract image features and construct a continuous homology distribution based on topological features for automatic tumor segmentation in histology whole-slide images. To achieve good segmentation results, some studies take CNN classification results as the initial segmentation value and improve the cell nucleus segmentation results using a level-set model [220] or sparse shape model [260].

(2) Lesion and tumor segmentation

Prior to treating lesions or tumors, the key step is to accurately segment the lesion or tumor to ensure that tumor cells can be killed and normal tissues or organs can be protected during treatment [54]. To segment the lesion and tumor accurately, multi-modal image information and global and local context information are usually combined. Therefore, some studies have employed multi-modal image information as inputs, with multi-processing flow networks being adopted for different image scales [84], and 3D CNN [59] and nonuniform sampling block strategies [84] [18] being utilized for segmentation tasks.

For comparison, we have combined several excellent representative algorithms that have been validated by the public dataset of the Brain tumor image segmentation bench-

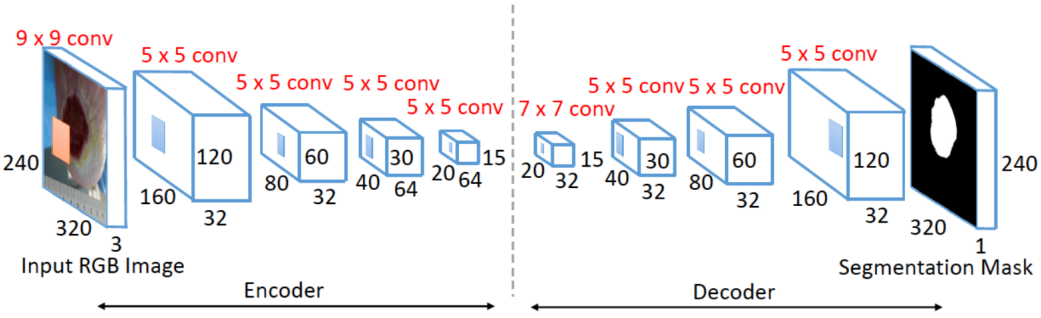
Table 2.2: Comparison of methods for brain tumor segmentation (validation on BRATS database)

Reference	Method	DICE		
		Total tumor	Core tumor	Active tumor
Expert evaluation	Medical training and experience	0.88	0.93	0.74
Urban [241]	Muiltimodal input, training with 3D CNN	0.87	0.77	0.73
Zikic [281]	The 3D cube image block is transformed into 2D image block, and training the 2D CNN network.	0.837	0.736	0.69
Havaei [84]	2D multi-modal input, dual path cascade CNN architecture, integrated local details and global information.	0.88	0.79	0.73
Pereira [181]	3×3 small convolution kernels, more CNN layers and nonlinear operation, and less filter weights.	0.88	0.83	0.77
Kamnitsas [109]	A dual path network framework witusing 3D CNN with 11 layers and small filters.	0.898	0.75	0.721

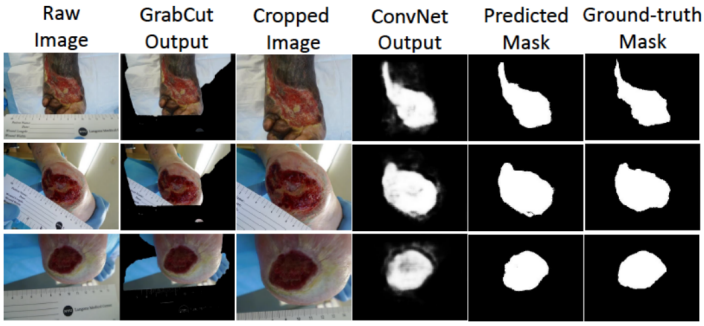
mark (BRATS)[5] over the past three years, as shown in Table 2.2. These algorithms are designed based on CNNs. Kamnitsas et al. [109] proposed the use of 3D CNNs with a dual pathway architecture to process input images with multiple scales simultaneously. This architecture incorporates both local and global contextual information. Three challenging lesion segmentation tasks, including traumatic brain injuries, brain tumors, and ischemic stroke, were evaluated using multi-channel MRI patient data. This method has demonstrated excellent performance beyond expert delineation level. Yu et al. [267] constructed a fully-convolutional residual network using a residual network and a full convolution neural network, which automatically segmented melanoma in dermoscopic images and won second place in the ISBI 2016 challenge.

In lesion segmentation tasks, we also observe the application of U-net, as well as the global and local similar framework. For example, Wang et al. [248] used a U-net structure with the same downsampling and upsampling paths; however, no jump connections were used in the network (as shown in Figure 2.10). Another framework similar to U-net was adopted to segment multiple sclerosis lesions using 3D convolution, and there was a single

[5]https://www.smir.ch/



(a) ConvNet architecture



(b) Segmentation results

Figure 2.10: ConvNet architecture for automatic wound segmentation results. (a) End-to-end approach for wound segmentation. (b) Wound regions cropped from raw images by modified GrabCut [199]. The cropped images are used as inputs and pixel-wise probabilities of the wound segment masks are taken as outputs (lighter means higher probability). A threshold of 0.5 is set to obtain the final masks on each pixel. This figure has been adapted from [248].

jump between the first convolution layer and the last deconvolution layer [18].

Note that most pixels in the image belong to normal tissues. One major challenge in lesion segmentation is imbalance in the class distribution. The coping strategies for class distribution imbalance are discussed in the section 2.4.

2.5 Analysis and Interpretation of Cardiac MRI Data

In clinical cardiology, cardiac function analysis plays an important role in patient management, disease diagnosis, risk assessment and treatment decision-making. Evaluating a set of complementary indices calculated from different structures of the heart using digital images is a routine task in cardiac diagnosis. Because CMR, which is constructed from the SAX view, has the ability to recognize different types of tissues, it is considered as the

golden standard of cardiac function analysis and helped evaluate the LV/RV Ejection fraction (EF) and SV, LV mass and myocardial thickness. This requires accurate delineation of LV endocardium and epicardium, as well as ED and ES conditions of the RV endocardium. In clinical practice, because the automatic cardiac segmentation method lacks accuracy, semiautomatic cardiac segmentation is still a standard practice. However, this can be time-consuming, easily lead to differences within and across observers.

Several difficulties have been identified in CMR segmentation. First is the poor contrast between the myocardium and surrounding structures and the high contrast between blood and the myocardium. Second, because of the blood flow, brightness heterogeneities existed in LV/RV chambers. Third, the intensities between trabeculae and papillary muscles are similar to the myocardium. In addition, nonhomogeneous partial volume effects arise because of limited CMR resolution along the long-axis, and there is inherent noise due to motion artifacts and heart dynamics. Finally, banding artifact exist.

In 2015, an estimated 17.7 million people died of CVDs, accounting for 31% of all deaths worldwide [176]. More people die from CVDs each year than because of any other cause. Clinicians have always relied on manual methods of tracking ventricular contours to obtain quantitative measurements, such as volume and mass. Generally speaking, a trained expert needs 20 min to analyze the images of a single subject at two time points of the cardiac cycle: ED and ES. This process is time consuming, tedious, and prone to subjective errors. Advances in medical imaging technology have led to a variety of non-invasive research options for CVDs, including echocardiography, CT and CMR. Each of these techniques has its advantages and disadvantages. Because of its good image quality, good soft tissue contrast and non-ionizing radiation. CMR has established itself as the non-invasive gold standard for evaluating the volume and quality of various CVDs [194][63][158].

Machine learning algorithms, especially DL networks, have shown great potential for many visual tasks. They can achieve or surpass human performance in many applications, including object recognition in natural images [86], game playing [217], tumor classification [61], and ocular image analysis [143]. In recent decades, DL based approaches have been applied in CMR image analysis [10][169][238][132]. Because of the limited size of the dataset, the majority of these works have used neural networks with relatively shallow architectures. In 2016, Kaggle provided 700 subjects for the second Data Science Bowl. In this challenge, all the data had no annotation [160]. Another challenge was organized by MICCAI, which provided 100 subjects with manual annotation [15]. Lieman Sifry et al. have compiled a dataset comprising 1143 SAX image scans [132]. Most of these images are labeled with LV and RV endocardium contours, with only 22% of them being labeled with LV epicardium contours.

2.6 Challenges in Deep Learning and Research Directions

2.6.1 Challenges in Deep Learning for Medical Image Analysis

DL is a data-driven approach for learning abstract features at all levels. DL demonstrates a very strong representation ability and robustness in many applications. Although DL demonstrates excellent performance in computer vision tasks where natural images are analyzed and processed, applying DL to medical image analysis is challenging. Various challenges are summarized as follows.

1. Natural images have higher spatial resolution and contrast than most medical images; furthermore, natural images have many visual features, such as brightness, color, and texture. Most medical images only have the intensity of a particular signal and a very low signal-to-noise ratio. Therefore, the boundaries between anatomical structures of organs and lesion areas in most medical images are unclear, with the texture differences not being obvious. Simultaneously, medical images differ significantly from natural images due to imaging principles. Thus, medical image analysis is more difficult than natural image analysis.

2. There are limitations in various medical imaging methods. Medical images obtained using different modalities can only provide specific anatomical and functional information of the human body, each having unique advantages and disadvantages. Different imaging devices and image reconstruction methods differ significantly, and different imaging principles and methods are typically used in clinical practice. Thus, automatic medical image analysis is more complicated than natural image processing.

3. Currently, many computer vision classification tasks are image-level tasks; however, medical images are used for image-level disease screening and pixel and voxel-level treatment planning. For example, intensity modulated radiation therapy (IMRT) requires accurate detection, identification, and localization of tumors, dangerous tissues, and organs. IMRT also requires high-precision segmentation of tumor areas and their surrounding normal tissues or organs from CT, MRI, PET, and other medical images. Moreover, the abnormal lesion area (e.g., tumor) is very complex, and the locations, sizes, and shapes of abnormal lesion areas vary greatly. Therefore, detection, recognition, and segmentation of an abnormal lesion area are more challenging compared with those of normal tissues and organs. The computation for medical image analysis is more complex than that for natural image analysis; thus, many DL algorithms in computer vision cannot be applied directly to medical image analysis.

4. In image classification tasks in computer vision applications, industry has established large-scale training datasets with manual labeling, such as the MNIST, CIFAR, and ImageNet datasets. However, it is very difficult to obtain a large training dataset when applying DL models to medical image analysis, especially for a lesion sample dataset because it varies a lot and requires clinical experts for annotation; thus, the available labeled data is limited [255] [254] [73]. However, detection, identification and segmentation of abnormal lesions are important in clinical applications, such as automatic screening, automatic diagnosis, and automatic treatment planning.

5. It is difficult to construct large annotated datasets for medical data analysis. For example, it is difficult to obtain financial support to build such datasets, and highly paid medical experts are required to annotate high-quality medical image data. Medical images are primarily located in private databases in hospitals, and privacy regulations may hinder access to such data, making sharing of medical data more difficult than sharing of natural images [82]. Medical image analysis is widely used, and many different datasets corresponding to different applications are required.

2.6.2 Coping Strategies

To improve feature representation and classification accuracy, large medical image training datasets are needed. However, several challenges must be considered. How do we deal with the shortage of training data? How do we use small amounts of training data in the most effective manner? How do we improve medical image classification accuracy using complementary information and image spatial context information? How do we obtain and annotate large medical image datasets? The current strategies for addressing such questions are summarized as follows.

(1) Transfer learning and weakly-supervised learning

The primary potential of a CNN lies in its ability to extract a series of discriminative features from multi-layer neural networks. As mentioned previously, a CNN is a supervised learning model, and training a CNN from scratch is a significant challenge. To address this issue, CNN models are typically pretrained in a supervised manner using natural images or datasets from different medical fields using the transfer learning method. There are two typical transfer learning strategies.

- Using pretrained network as feature extractor. A CNN model trained by ImageNet can be used in medical image recognition despite the differences in imaging principles and the appearance between medical and natural images [22]. For example, Bar

et al. [13] [14] used a pretrained network as a feature extractor for chest pathology recognition, and Ginneken et al. [243] combined CNN features with handcrafted features to improve nodule detection performance.

- Using target medical data to fine tune the pretrained network. A pretrained CNN has been used as the initial network, with the network parameters fine-tuned via a supervised approach with limited annotated data to adjust parameters in several or all network layers [91] [28]. Ciompi et al. [39] proposed to automatically detect pulmonary fissure nodules using a CNN pretrained by ImageNET and fine-tuning the network with a small number of labeled CT data sequences. Tajbakhsh [231] demonstrated that the performance of deep fine-tuning is better than that of shallow fine-tuning and that the importance of fine-tuning a network is enhanced when the size of the training dataset is reduced.

Both strategies have been widely used. The first strategy has an additional advantage, *i.e.*, it does not require training the deep network and allows extracted features to be inserted easily into existing image analysis pipelines. However, few studies have thoroughly investigated which of the two strategies yields better results [136].

In addition to transfer learning, another strategy is weakly-supervised learning, which effectively combines the advantages of unsupervised and supervised learning. Although the number of medical images with available annotations is small, the unlabeled data scale may be large. We can make full use of unlabeled data and adopt pretraining technology to extract features for images in an unsupervised manner. Then the extracted features can be used as the initial value for a supervised learning network, and the classifier can be trained using the limited labeled data. For example, Kallenberg et al. [107] used a CSAE for breast density segmentation and breast risk assessment. The main difference between the CSAE and a classical CNN is that the convolution layers of the CSAE are pretrained layer by layer in an unsupervised manner as an SAE. The combination of multiple instance learning (MIL) and DL is also a feasible alternative in case the acquisition of annotated data is very expensive. Xu et al. [263] studied MIL frameworks combining supervised and unsupervised feature learning approaches. The results demonstrated that the performance of MIL-based frameworks is superior to that of handcrafted features and close to that of supervised methods.

(2) Regularization and equalization

CNN training is an iterative process of parameter optimization. In each iteration, a sample is selected randomly from the training data as the input to the network, and the

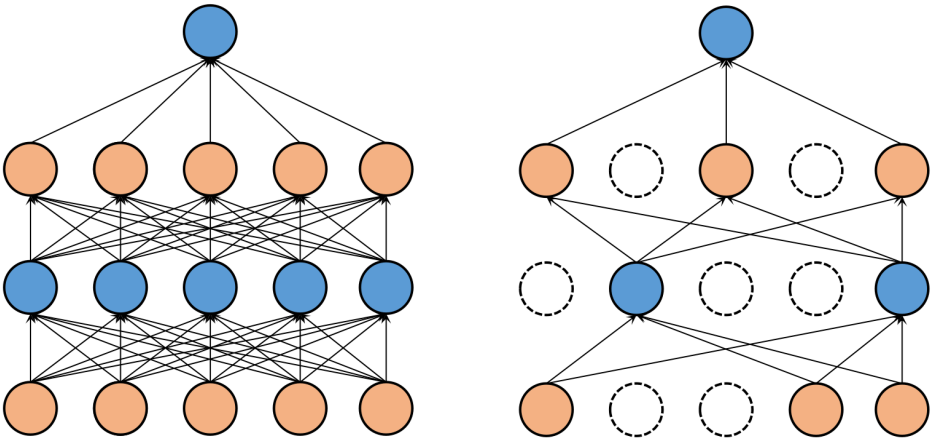


Figure 2.11: Dropout neural net model. *Left*: standard two-layer neural network. *Right*: dropout is applied in the standard network.

parameters are updated by backpropagation to minimize the objective function. The differences in medical imaging equipment and image reconstruction methods may cause uneven gray level and inconsistent offset field problems [276]. Although supervised learning techniques have demonstrated great potential in experiments using standardized imaging protocols, performance may deteriorate rapidly with new input images captured under slightly different conditions. In light of such issue, we should consider the following processes to weaken the adverse effects.

- Batch normalization: By normalizing the mean and variance of each training sample, we can avoid gradient disappearance, gradient overfitting, and accelerate convergence. This process can be used as a regularization technique to improve the network's generalization ability [99].

- Regularization: Regularization is an effective strategy to reduce overfitting. By adding regularization terms L1 or L2 to the model's cost function, the complexity of the model is reduced, which, in turn, reduces overfitting [84].

- Dropout: In each iteration, the output of partial neurons whose proportion is p is randomly set to 0 (*i.e.*, some nodes are disconnected). Dropout is a random regularization strategy to avoid overfitting of the network, and it can be considered for implicit model integration [89], as shown in Figure 2.11.

Classification tasks in medical image analysis typically need to distinguish between the normal tissue and lesion area. In each case, the data distribution of various tissue types is unbalanced, and most normal tissue and organ samples are highly correlated and possess

a large amount of information; thus, the normal tissue and organ can get overrepresented. For example, brain tumor segmentation is a highly data-unbalanced problem, in which the voxels of healthy tissue account for approximately 98% of the total voxels, and the remaining 2% voxels of pathological tissue include 0.18% belonging to necrosis, 1.1% belonging to edema, 0.12% belonging to non-enhanced tumors, and 0.38% belonging to enhanced tumors [84]. Treating these data equally in the learning process will result in many iterations wasted on non-information samples, thereby rendering the training process unnecessary. Simultaneously, this type of training, *i.e.*, training dominated by healthy tissue samples will lead to problems in the CNN model. To address this problem, van Grinsven et al. [244] improved the learning efficiency of a CNN and reduced training time by identifying normal samples with large amounts of information and dynamically selecting negative samples with misclassification in the training process. Havaei et al. [84] proposed a two-time training strategy, *i.e.*, initially selecting all kinds of images with equal probability for training, and then keeping all layers' kernels fixed using the more representative true distribution of the samples to retrain the output layer. As a result, the diversity of all classes can be balanced, and the output probability can be corrected by retraining the true distribution of the data labels. Brosch et al. [18] adjusted the loss function and defined it by combining weighted sensitivity and specificity. The greater the specific weight, the less sensitive it is to data heterogeneity. There are also ways to balance data distribution via data augmentation of the positive samples [109] [181].

(3) Integrating multi-modal complementary image information and image spatial context information

Due to the limitations of medical imaging, medical image data of different modalities can only reflect the specific information about the human body, and each modality has its advantages and disadvantages. For example, CT and MRI complement each other in bone and soft-tissue imaging, and CT, MRI, and PET complement each other in anatomical and functional imaging. Different contrast enhancement methods offer different advantages even when used with the same imaging method. For example, different MRI modalities produce images with different tissue contrasts, thereby providing valuable structural information and enabling diagnosis and segmentation of tumors and their clinical regions. The complementary information contained in multi-modality medical images can provide clear functional and anatomical structure information, which improves the accuracy of analysis. Therefore, most segmentation algorithms use multi-modality medical images as inputs [278] [224].

To achieve pixel-level medical image classification/segmentation, classifying only the pixels is insufficient; thus, we must also combine neighborhood pixels to provide better context information. Most medical images provide 3D information. To better consider voxel context information, in addition to adopting a deeper CNN model, using multi-scale and multi-processing flow CNN methods, we must also consider how to model 3D information. Currently, there are two different ways to handle 3D information modeling.

- Transforming 3D information into 2D image block information. Taking the classified voxels as the block center, the 3D neighborhood context information is expressed approximately using multi-view 2D profiles, and the computation is simplified via 2D convolution; thus, computational efficiency is improved [197] [281]. By considering each section as an information source, all section information can be integrated via a multi-channel or multi-processing flow to improve classification accuracy. For example, Roth et al. [197] used three individual and separately trained CNNs on each orthogonal image slice, with a subsequent fusion of their predictions to detect colonic polyps or suspicious lymph nodes. Setio et al. [211] used a CNN-based multiprocessing flow to classify the points of interest of chest CT images. By extracting the features from nine different directions of the points of interest as the input, these features were merged at the fully-connected layer to obtain the final classification results.

- Processing with a 3D CNN. Taking classified voxels as the block center, 3D cube blocks are extracted and processed via 3D convolution. This method considers the 3D neighborhood information comprehensively and helps extract more discriminative features; thus, classification accuracy is increased. This method has a disadvantage that too many 3D elements may be involved in the computation and computational efficiency is low [109]. With increased computing speed, 3D CNNs have been widely adopted in the past couple of years. For example, Nie et al. [171] trained a 3D CNN using 3D information to evaluate the survival rate of patients with severe glioma.

In addition, we can improve the accuracy of the classification and refine the boundary of the region for segmentation by combing different algorithms. For example, we can use the super-pixel segmentation method to extract the candidate regions of interests. Then deep learning is used to extract deep features for these regions. This method can reduce the search space to improve computational efficiency and classification/segmentation accuracy [221] [197] [196]. Ngo et al. [169] proposed a method for automatic and accurate segmentation of LV from cardiac MRI using DL and level-set method .

2.6.3 Open Research Directions

In summary, DL models can automatically learn more discriminative features from data. DL models have been applied to many medical image analysis tasks and they have helped achieve significant breakthroughs. In most studies, DL methods have been used to demonstrate their leading performance, such as successful application of several computational challenges in medical image analysis. In addition, with the development of cloud computing and high-performance multi-graphics processing unit parallel computing, it is possible to learn deep features from massive medical image data. Finally, the emergence of publicly accessible medical image databases, such as the brain tumor MRI dataset (BRATS), the Alzheimer's disease neuroimaging dataset (ADNI), the ischemic stroke dataset (ISLES), and various medical image segmentation challenge datasets, has facilitated effective validation of DL-based segmentation algorithms [214], [200].

Most of the advanced DL methods are supervised learning approaches, specifically CNN-based frameworks. Previous studies have focused on pretrained CNNs and using CNNs as a feature extractor that can be downloaded easily and used directly for medical image analysis. End-to-end training for CNNs has become a priority in medical image analysis. However, obtaining annotated data for supervised learning is a significant challenge compared with applying DL methods in medical data analysis [254] [200]. Under the condition of limited labeled training data, it is important to make full use of non-labeled images in medical image analysis. In addition, it is expected that weakly supervised learning methods that combine the advantages of unsupervised and supervise learning will yield practical benefits.

The text reports of medical experts and electronic medical records contain rich clinical information, which can be used to supplement labeled image data. In the computer vision field, it is expected that natural image subtitle generation methods combined with RNNs and CNNs will soon be applied to medical image analysis.

These challenges provide tremendous opportunities to medical image analysis researchers. We believe that through improvements in DL algorithms, the development of high-performance parallel computing technology, the increasing quality of medical images, and the growing amount of labeled medical image data, DL-based medical image analysis will achieve great success in the future.

2.7 Quantitative CMR Image Analysis of UK Biobank

Quantitative assessment of cardiac function is essential for appropriate preventive care and early CVD treatment. In large-scale population imaging data, the analysis and in-

terpretation of cardiac structural and functional indicators can help identify patterns and trends in different population groups and thus reveal the key risk factors of CVD before comprehensive CVD development. UK Biobank (UKBB) is one of the largest prospective population studies worldwide, which aims to investigate the determinants of a disease [183]. UKBB data include a wide range of baseline questionnaire data, biological samples, physical measurements and CMR images to establish cardiovascular imaging derived phenotypes [182]. In many UKBB imaging centers, CMR is an important part of multi-organ and multi-modality imaging visits for patients. The centers will acquire and store imaging data from 100,000 participants by 2022.

Cardiac ventricle segmentation has always been a hot topic in the field of medical image processing. The purpose of cardiac ventricle segmentation is to derive quantitative measurements of cardiac ventricles, such as LV EF, volume and mass, which can be calculated from segmentation results to evaluate cardiac function. Image segmentation is the process of dividing an image into several specific, unique regions and processing the object of interest; it is the key step from image processing to image analysis. Automatic segmentation is designed to reduce the tedious, time-consuming and error-prone tasks. So far, many related algorithms have been developed, ranging from the most basic region partitioning technology and graph-based segmentation algorithms to machine learning [47] and CNN based deep learning (DL) algorithms [12]. We will give a detailed literature review on DL based medical image segmentation methods in section 2.3.3.

2.8 Thesis Overview

In the previous chapter, we reviewed background information regarding cardiac image analysis problems and state-of-the-art analysis methods focusing on DL. In particular, we discussed some proposed methods that involve a CNN for cardiac image segmentation, multi-modal registration problems. Chapter 3 introduces the intensity representation learned by a CNN for LV coverage assessment in the UKBB. In Chapter 4, we extend a previously proposed 2D CNN to a 3D CNN and introduce the Fisher-discriminative (FD) criterion. In addition, we experimentally demonstrate the applicability of the FD criterion in a framework with a small amount of training data. In Chapter 5, we explore the concept of AL based LV and RV coverage assessment and experimentally demonstrate its benefits on LV+RV coverage assessment for datasets with no label information. In Chapter 6, we propose a strategy for CMR plane pose estimation and a multitask regression model with PI, which encourages CNN models to make more meaningful pose predictions automatically. Then, in Chapter 7, we propose using quality aware GANs for CMR images to

synthetize the missing slice. Finally, in Chapter 8, we summarize the work presented in this thesis and discuss future work.

Chapter 3

Automated LV Coverage Assessment for Cardiac MR Images Using Convolutional Neural Networks

This chapter is based on:

- **Le Zhang**, Ali Gooya, and Alejandro F. Frangi, Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets, *MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 61-68. Springer, Cham, 2017*

- **Le Zhang**, Ali Gooya, Bo Dong, Rui Hua, Steffen E. Petersen, Pau Medrano-Gracia, and Alejandro F. Frangi, Automated quality assessment of cardiac MR images using convolutional neural networks, *MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 138-145. Springer, Cham, 2016*

Authors' contributions: L.Z., A.G., B.D. and A.F.F. conceived and designed the study; S.E.P. provided support on clinical aspects and he also provided the UK Biobank data resource to be used for training and testing; R.H. provided the image mask for the data. L.Z. designed the method, performed data analysis and wrote the manuscript. All authors read and approved the manuscript.

3.1 Introduction

In this chapter, we present a new problem in medical images analysis, namely image quality assessment (IQA), and its specific problem in cardiac MR image analysis tasks such as LV coverage assessment. In video processing, Automatic Image Quality Assessment (AIQA) is a well-developed corpus of techniques usually concerned with detecting image distortions characteristic of multimedia communications [202] [264]. These distortions are generally very different to those affecting medical imagery. No-reference based image quality assessment (NR-IQA) [87], [164] is relevant for medical imaging data since while it is easy to get access to abundant data sets of mixed quality, it is infeasible to collect data without some level of image degradation or artifacts.

Visual quality is a very complex yet inherent characteristic of an image. In principle, it is the measure of the distortion compared with an ideal imaging model or perfect reference image. The characters of LV are useful to identify the position, which the slice belongs to, since the LV in each slice shows a different shape and size. Recent work [67], [259] has focused on learning data-driven features in order to more accurately detect shape differences. Among them, convolutional neural networks (CNNs) are one of the most regularly used deep learning schemes to meet the challenges of discriminative shape detection [237] [136].

Based on these observations, we explore using a CNN to learn discriminant features for the LV coverage assessment task for cardiac MR images. Recently, deep neural networks have gained researchers attention and achieved great success on various computer vision tasks. Specifically, CNN has shown superior performance on many standard object recognition benchmarks [111] [118] [41]. One of CNNs advantages is that it can take raw images as input and incorporate feature learning into the training process. With a deep structure, the CNN can effectively learn complicated mappings while requiring minimal domain knowledge.

We demonstrate two different methods of deep learning for LV coverage assessment in cardiac MR images: (1) 2D convolutional neural network, (2) Semi-coupled generative adversarial networks (SCGAN). All the two methods make use of the UK Biobank dataset by focusing on the analysis of short axis (SA) cine MRI. At the end of the chapter, we demonstrate the desighed framework can achieve effective generalization properties, when applied to complex classification problems such as identifying missing SA slices.

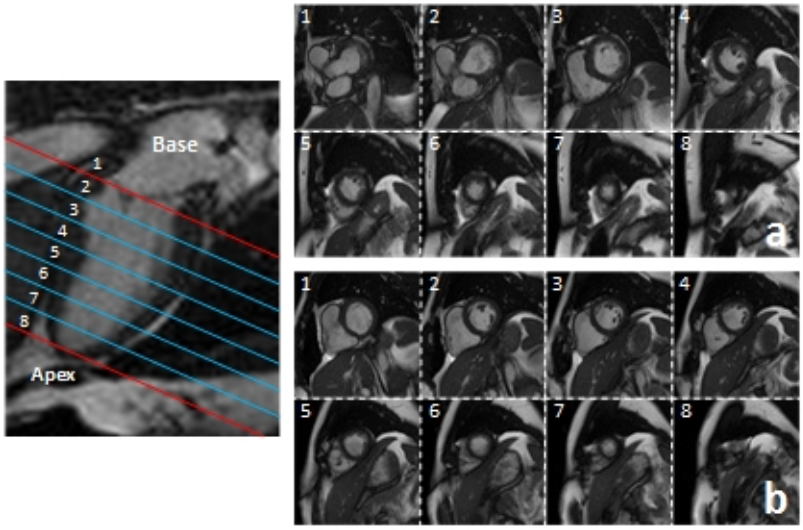


Figure 3.1: Left: A typical two-chamber view cardiac MRI with eight slices covering from base to apex; Right: (a) a volume with whole coverage (slice 1 is the basal slice), and (b) a volume with missing basal slice (slice 1 is notthe basal slice). In each rectangle, from top to bottom, rows correspond to adjacent axial slices.

3.2 Automated Quality Assessment of Cardiac MR Images Using Convolutional Neural Networks

Cardiac Magnetic Resonance Imaging (CMRI) can not only reflect anatomic information of the heart but also provide physiological information associated with cardiovascular diseases. Although low image quality can be minimized by careful design of the imaging acquisition protocols, it cannot be fully avoided; particularly in large-scale imaging studies, where data is acquired at different imaging sites, across subjects with a diverse constitution and at a big pace [62].

On the other hand, few objective guidelines exist, clinical or otherwise, that establish what constitutes, in general, a good image and, in particular, a good CMRI study [242]. To ensure that the quality of data collected in such imaging studies is maintained, Image Quality Assessment (IQA) is crucial. Surprisingly, IQA is still usually carried out by visual inspection of the images which can be exhaustive, costly, subjective, error prone, and time consuming [9]. Thus, Automatic IQA (AIQA) methods are required to detect deviations from the desired quality, intervene to correct problems in data collection as soon as possible, and discard low-quality images, whose analysis would otherwise impair any aggregated statistics over the cohort. Additionally, *a priori* and objective knowledge on image quality of a given dataset (and possibly the type of artifact affecting it) could assist

in choosing the most appropriate image analysis method to be used. This paves the way to "quality-aware image analysis" [253].

In multimedia, AIQA is a mature research field and usually concerned with detecting specific image distortions [202] [264]. Unfortunately, most of these methods cannot be directly translated to medical imaging due to different properties in image statistics and the more complex nature of image artifacts [119]. Thus, AIQA remains as a relatively unexplored research area in medical imaging. It is acknowledged that lack of basal and/or apical slices is probably the most common problem affecting image quality in CMRI and has a major impact on the accuracy of quantitative parameters of cardiac performance [114]. In this study, we mainly focus on short axis (SA) cine MRI. More specifically, we aim to identify missing apical slice (MAS) or missing basal slice (MBS). To address this problem, we are motivated by the success of deep learning techniques and, in particular, Convolutional Neural Network (CNN) [15][32]. They can achieve effective generalization properties, when applied to complex classification problems such identifying missing SA slices.

To the best of our knowledge, this is the first study tackling the problem of detecting the missing slices in CMRI. Apart from introducing a new application for the CNN's, and addressing a pressing need, we propose an effective strategy for their training. In practice, the lack of sufficient number of CMR data sets with MBS/MAS deficiencies imposes a severe class imbalance problem. To alleviate this issue, only the bottom and top SA slices are examined to ensure the full coverage of the heart. This allows us to use the middle slices as non BS/AS training samples. We present results for various depth of the networks, and identify the optimal number of the layers. We also compare our framework with an array of other deep learning methods such as Deep Boltzman Machines (DBM) and Stack Auto Encoders (SAE), and show its better performance. In the next section, we briefly introduce the architecture of our networks and provide the specification of our data sets. We then present our classification results and conclude the study in the final section.

3.2.1 Methodology

As mentioned, we are interested in detecting missing apical and basal slices in CMRI data sets. To this end, for each cardiac subject, the top and bottom SA slices in the scan are classified using two CNNs, each particularly trained for detecting missing slices in basal or apical positions. Each CNN is composed of alternating convolutional and sampling layers, and one fully-connected output layer. Figure 3.2 shows the configuration of CNNs with total number of 5 layers (showing overall the best classification performance). Here, we briefly review the various components in the proposed CNNs with a further detail.

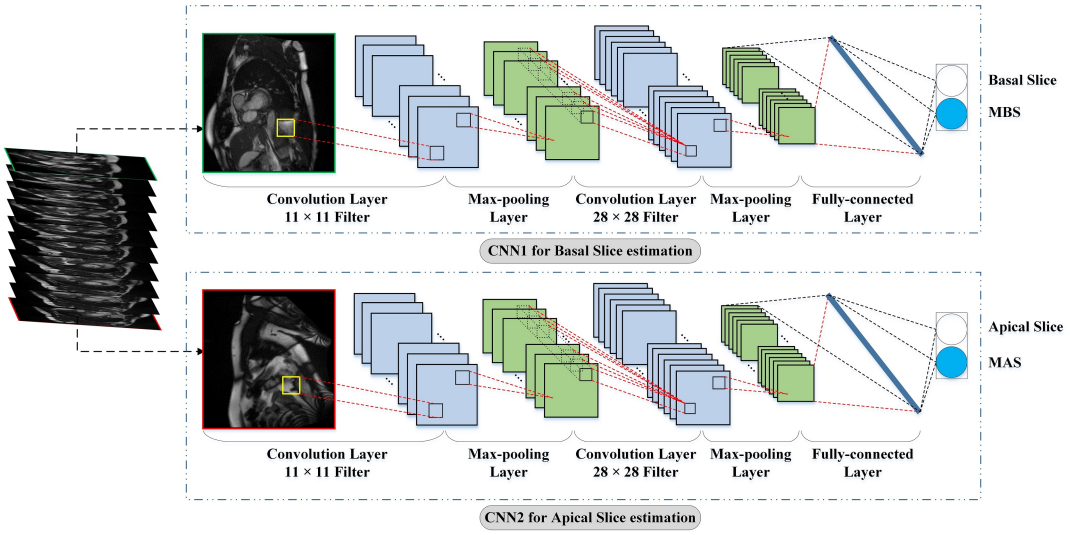


Figure 3.2: Overview of our proposed deep learning model for cardiac MRI quality assessment. The CNNs are composed of 5 layers: four multi perceptron convolutional layers plus one fully-connected layer. The bottom and top SA slices are examined individually.

Convolutional Feature Layers: Convolutional layers implement kernels that are used to detect discriminative features from input images [177]. During the training, these kernels are optimized to compute some salient features (such as edges, corners, etc.) that are relevant for discrimination of the observed categorical variables. We define $\mathbf{X}_i^{l-1}$ and $\mathbf{X}_i^l$ as input and output ith feature map of the lth layer. Let $m \times n$ and $k \times k$ be the size of input maps and the convolution kernel for layer l. With this setting of parameters, we can get N output maps with the size $(m-k+1) \times (n-k+1)$. The output of a convolutional layer l is given by

$$\mathbf{X}_j^l = f\left(\sum_{i \in M_j} \mathbf{X}_i^{l-1} * \mathbf{k}_{ij}^l + b_j^l\right), \tag{3.1}$$

where $\mathbf{k}_{ij}^l$ denotes the convolution kernel linking the ith input to the jth output map; b_j^l is the bias vector for the jth output-feature-map of lth layer; f is the activating function $1/(1+e^{-x})$, and M_j is the input feature map in the former layer.

Sampling Layers: These layers are designed to reduce the number of kernel parameters, minimize the computational complexity, and make the features robust to zoom, shift and rotation. The output of convolution layers are divided into sub-regions having the size of $w \times h$ pixels. Then, each output pixel of a sampling layer is defined as the maximum value in the corresponding input sub-region. These operations can be formulated using the

following relationship

$$\mathbf{X}_j^l = f\left(\beta_j^l down\left(\mathbf{X}_j^{l-1}\right) + b_j^l\right), \tag{3.2}$$

where $down(\cdot)$ symbolizes the down sampling function; j, l, β and b denote the feature map index, the layer number, the weighting coefficients, and the bias vector, respectively.

Softmax classifiers: To predict the final labels, the CNN detected low-dimensional features are used to train softmax classifiers. Given the feature vector $\mathbf{x}^{(i)}$, we computed the posterior probabilities for $k = 1, 2, ..., K$ classes using

$$p(y^{(i)} = k|\mathbf{x}^{(i)}) = \frac{e^{\boldsymbol{\theta}_j^T \mathbf{x}^{(i)}}}{\sum_{l=1}^{K} e^{\boldsymbol{\theta}_l^T \mathbf{x}^{(i)}}}, \tag{3.3}$$

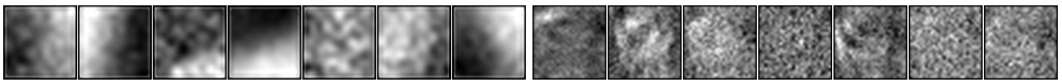
where $\boldsymbol{\theta}$ denotes the parameters of the softmax classifier, obtained from the pre-trained CNN network. The neural network was trained over 3 days for 100 epochs with a fixed learning rate 0.01. In the framwork, Rectified Linear Unit (ReLU) [118] was used as a activation function, and back-propagation technique [201] was used for adjusting weights of connections in the network. To test a single image with size 100×100, it only took approximate 0.2 seconds.

Combining outputs from the classifiers: As mentioned, we are interested in detecting missing apical and basal slices in CMRI data sets. In the second quality estimation step, the final Cardiac MRI subjects were classified into different classes using a logical classifier. Figure 3.2 illustrates the process of the classification of cardiac images subjects quality.

To this end, for each cardiac subject, the top and bottom SA slices in the scan are classified using two CNNs, each particularly trained for detecting missing slices in basal or apical positions. To obtain the final classification, we then combine the predicted outcomes of each classifier to receive the final quality category of each subject. For the two steps framework training, we firstly identify whether our input stacks miss the apical or the basal slice, then take outputs from each pre-trained CNN as inputs of the logical classifier, which combines the two CNN outputs to have the final image quality category.

3.2.2 Experiments and Results

Region of Interest Extraction: Since the heart is the object of interest for quality assessment and minimize the influence from the background region,, a mask covering the heart and its surrounding structures was globaly employed to remove unnecessary information in the background. The region of interest was extracted by detecting the area of cardiac motion as follows: the average absolute image intensity difference was computed



(a) (b)

Figure 3.3: The learned convolution kernels on basal and mid-slices of the first (a) and the second (b) layers of the trained CNN.

for the whole sequence and used for thresholding the image sequence producing a binary mask. It is assumed that the largest region in this binary image is relevant to the heart. as observed in the image. Thus, the binary mask can be further refined by removing regions which are far from the cardiac region. Finally, the ROI was obtained as the region covering the detected cardiac region.

CMRI Quality Criteria: In the apical slice, the LV cavity is still visible at end-systole, and the left ventricular outflow tract (LVOT) is existing in the basal slice [114]. The absence of basal slices has an important impact on the volume calculation, and missing middle slice and apical slice can result in a penalty as well. Thus, we define four classes of qualities in this study: MAS, MBS, missing apical and basal slices (MABS), and no missing (normal). The last label is obtained by logical combination of the results from the MAS and MBS classifiers.

Table 3.1: The average precision and recall rates of each type of missing slices using different deep learning models.

	Precision Rate			**Recall Rate**		
	MAS	**MBS**	**Normal**	**MAS**	**MBS**	**Normal**
SAE	79.08%	68.63%	78.54%	88.48%	88.72%	88.15%
DBM	66.67%	70.09%	71.47%	88.38%	88.71%	88.32%
3-CNNs	80.77%	70.92%	78.43%	88.52%	88.75%	87.85%
5-CNNs	**81.61%**	**74.10%**	**79.42%**	**88.73%**	**88.75%**	**88.01%**
7-CNNs	82.19%	69.43%	75.06%	88.62%	88.76%	87.01%

We apply our framework to 100 UK Biobank (UKBB) cardiac MRI pilot data sets. These data sets are obtained by 1.5T MR scanners [183][182] and show overall good quality and no missing slices. Therefore, to generate synthetic deficiencies in the data, we manually removed basal slices from 50 subjects and apical slices from another 50 subjects. For each kind of the considered defect, we randomly selected 80% of generated data sets as training

sets and the left the rest as the testing sets. In order to evaluate our proposed framework's performance, we use

$$Precision\ Rate = \frac{TP}{TP+FP}, \tag{3.4}$$

$$Recall\ Rate = \frac{TP}{TP+FN}, \tag{3.5}$$

where TP, FP, and FN are the numbers of the true positive, false positive, and false negative samples, respectively.

Evaluation and Comparison to other Deep Learning Models We systematically compared our proposed CNNs framework with different types of CNNs architectures and traditional deep learning methods. Table 3.1 lists the results for different CNNs architectures and other state-of-the-art deep learning methods. As seen, the CNNs with a total number of 5 layers shows the best precision rate and recall rates.

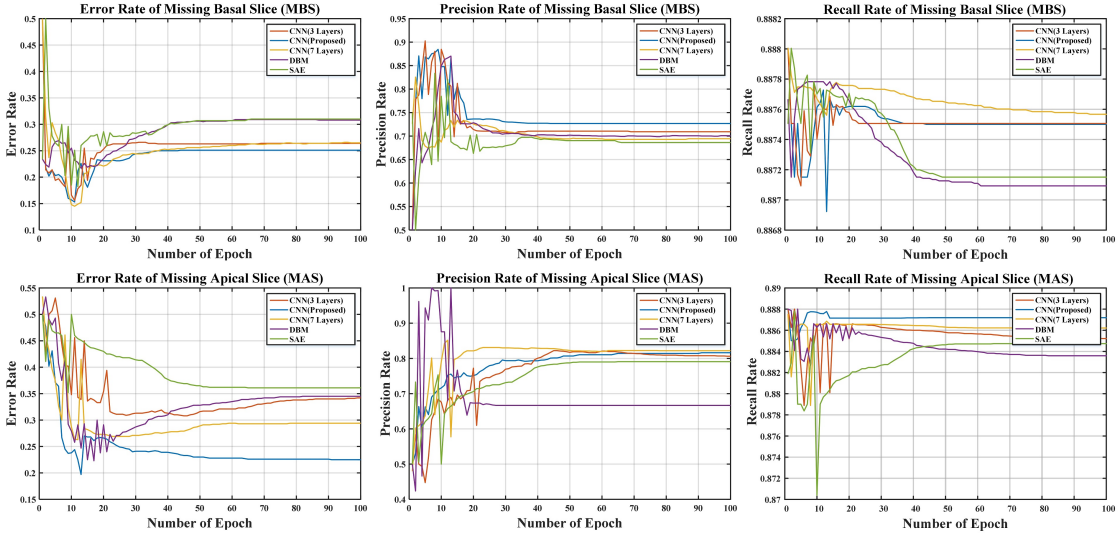


Figure 3.4: The distributions of the error, precision, and recall rates over 100 training epochs, showing a superior performance of the CNNs with 5 layers.

We also visually examined the learned convolution kernels, and found only a few kernels present structure related appearances. Figure 3.3 shows the kernels learned for classifying missing basal slices. It is not surprising that some of these kernels show noisy, rather than strong structural and interpretable patterns. This is because our features are trained to be discriminative. In fact, to obtain user interpretable features, generative models such as those outlined in [125] is usually considered.

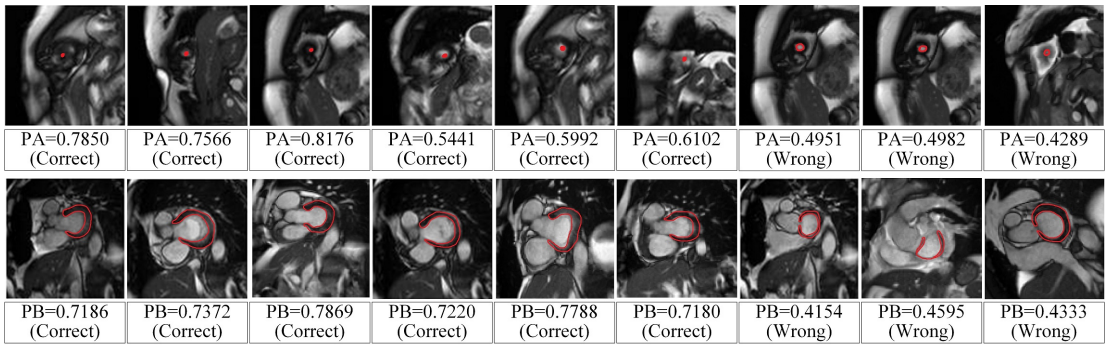


Figure 3.5: Sample test slices and their probability values of being apical (top row) or basal slice (bottom row) are shown. 'PA' means the Probability value of being Apical slice; 'PB' means the Probability value of being Basal slice. The 'correct' and 'wrong' subscripts indicates the classification results. Red segmentation shows the difference of ventricular contour for each subject.

Furthermore, to demonstrate the convergence behaviour of the compared methods, in Figure 3.4 we show the distributions of the error, precision, and recall rates over 100 training epochs. It can be seen the CNNs with 5 layers outperforms other CNN architectures and learning models.

In Figure 3.5, a few apical (top row) and basal (bottom row) slices in the test datasets along with their corresponding posterior probability values are shown. We can observe that our framework correctly classifies a few challenging basal slices, but also fails in a few other cases. Furthermore, the basal slices with existing LVOT's indicate higher probability values of being correctly classified. This shows that the training has been successful in capturing the LVOT as a prominent feature in the correctly positioned basal slices.

We also designed a validation experiment with a second collection of CMR data sets to show the generalization ability of our method. To this end, we trained the proposed model using the UK Biobank datasets and tested it using the data sets available from Data Science Bowl Cardiac Challenge data sets [17]. This experiment was repeated for 100 training epochs and the values for error, precision and recall rates are shown in Figure 3.6. These results show that our trained convolutional neural network achieves a good generalization efficacy.

3.2.3 Conclusion

In this study, we tackled the problem of identifying the missing apical and basal slices in large imaging databases. We illustrated the concept by applying the method to CMRI studies from the UK Biobank pilot datasets. We designed slice classifiers and learned a set of discriminative features directly by training Convolutional Neural Networks. Casting

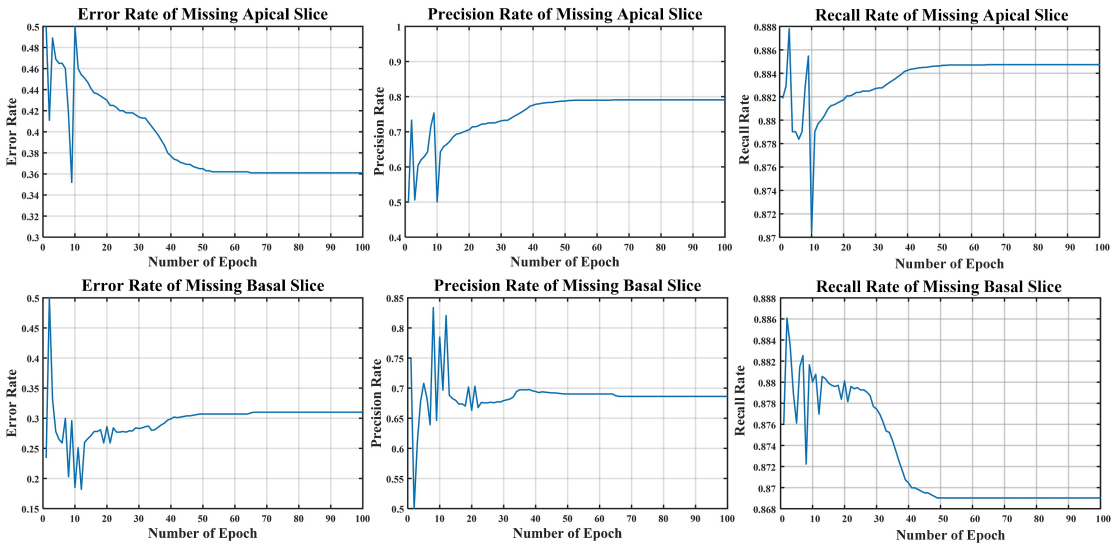


Figure 3.6: The error, precision, and recall rates in cross dataset test.

this problem as a slice classification task, we were able to alleviate the class imbalance issue and effectively train the CNNs using the available data. Different numbers of network layers were examined and compared to other deep learning models (such as Stacked Auto-Encoder and Deep Boltzmann Machines). We showed that a CNN model with 5 layers outperforms the other models. We also validated our model by training the 5-CNNs using UKB pilot datasets and applying them to CMR data sets from Data Science Bowl Cardiac Challenge. The proposed model shows a high consistency with human perception and becomes superior compared to the state-of-the-art methods, showing its high potential. In this study, the kernel sizes in the convolutional layers of the network were selected somehow arbitrarily. However, in principle these parameters can be optimized by performing exhaustive cross validation experiments. In future, we will further refine the current structure of our model by tuning such parameters.

3.3 Semi-supervised Assessment of Incomplete LV Coverage in Cardiac MRI Using Generative Adversarial Nets

In medical imaging it is hard to have access to quality-labelled image databases due to the diversity of image characteristics, and their artifacts, of diverse anatomical locations and image modalities. Therefore, it is essential to devise techniques that do not require manual labelling of visual image quality. Image synthesis models provide a unique opportunity for performing unsupervised learning. These models build a rich prior over natural image statistics that can be leveraged by classifiers to improve predictions on datasets for which few labels exist [174]. Among them, generative adversarial networks (GAN) can synthesize adversarial examples, which increase the loss by a machine learning model [230]. Meanwhile, GAN can perform unsupervised learning by simply ignoring the component of the loss arising from class labels when a label is unavailable for a training image [81].

In this study, we mainly focus on the analysis of short axis (SA) cine MRI. We aim to identify missing apical slices (MAS) and/or basal slices (MBS) in cardiac MRI volumes. In previous research, Le [271] used convolutional neural network (CNN) constructed on single-slice images and processed them sequentially. But this solution needs large amount of labelled data and lacks the ability to classify examples with perturbations correctly. In this study, we exploit semi-coupled-GANs (SCGANs), a semi-supervised approach, for incomplete LV coverage detection. To alleviate the lack of sufficient numbers of CMR datasets with MBS or MAS, the proposed SCGANs use two generative models to synthesize adversarial examples. By learning adversarial examples, it improves not only robustness to adversarial examples, but also generalization performance for original examples. This work is the first work we know of to use adversarial examples to improve the robustness of an attribute learning model.

3.3.1 Methodology

We present a novel technique of LV coverage assessment for CMRI by using SCGANs. The motivation behind our proposed method is: In medical image quality assessment problems, we are always faced with a lack of quality-labelled data, especially images with artifacts. Several deep learning models cannot classify the examples with perturbation correctly. Our semi-supervised SCGANs is proposed by using adversarial examples as the outlying observations for discriminative model training. We generate adversarial samples

by two generators separately, which confuse the discriminator into mistaking them for genuine images. After that, we obtain the robust attribute classifiers by learning both original data and synthetic data. Our proposed SCGANs represents a strategy to better handle the typical LV coverage assessment problem.

3.3.1.1 Generative Adversarial Learning

Recently, GAN [81] was proposed as a novel way for adversarial learning. It consists of a generative model and a discriminative model, both are realized as multilayer perceptrons [140]. The aim of the discriminator is to correctly classify the original examples and adversarial examples. By learning the adversarial examples, the network cannot only becomes robust to adversarial examples, but also generalization improves for unmodified examples. GAN does not need the label information when training the generator and then the discriminator can estimates the probability that a sample came from the original data rather than the generator.

We assume a probability distribution M, which is a black box relative to us. To realize how the black box works, we construct two 'adversarial' models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample from the training data rather than G. Both G and D could be a non-liner mapping function, such as a multi-layer perceptron. Our objective is to learn feature representation to handle a wide range of visual appearances in cardiac MRI and identify images with incomplete LV coverage. We regard adversarial examples as outlying observations regarding other samples in training data. The generative model constantly produce new adversarial samples and the discriminative model classify the positive and negative samples by learning the new produced adversarial samples constantly. Given a particular describable visual attribute - say 'MBS'. An outlier image is expected to be mapped to negative values, which indicates the absence of basal slice. This can happen for two reasons: (1) the image does not belong to the basal slice, (2) the image belongs to the adversarial examples. We consider them all as the outliers.

3.3.1.2 Semi-Coupled GANs

Here we introduce our model based on the above discussion. Our model is illustrated in Figure 3.7 designed as a semi-coupled-GANs for attribute learning. It consists of a pair of $Generators-$ G_1 and G_2, which share a same discriminator. Each generator synthesizes the adversarial samples Y_1 and Y_2 for positive and negative data, respectively.

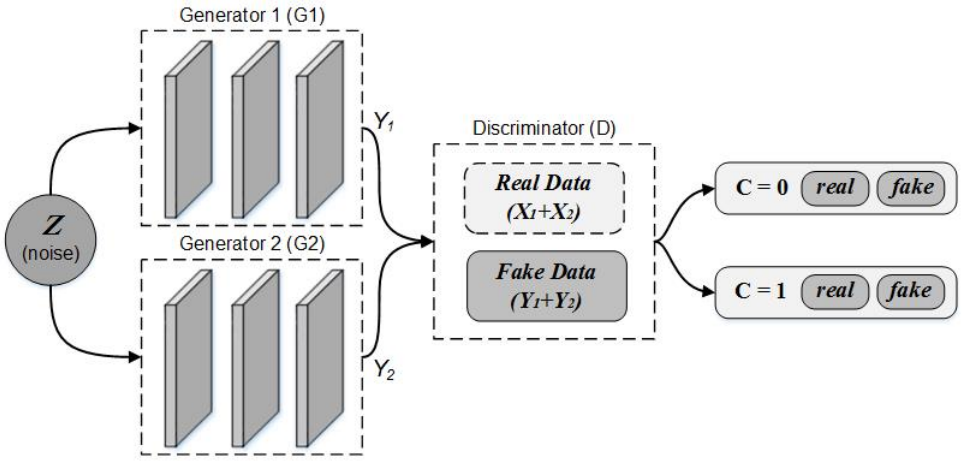


Figure 3.7: The Proposed Semi-Coupled-GANs Framework.

Generative Models: We firstly feed the two generators G_1 and G_2 noise data z, G_1 and G_2 learn probability distribution from the original positive and negative images respectively, and generate the corresponding adversarial samples. Then, we give the adversarial data to discriminator D. Denote the distributions of $G_1(z)$ and $G_2(z)$ by p_{G_1} and p_{G_2}. Both G_1 and G_2 are realized as multilayer perceptions:

$$\begin{cases} G_1(z) = G_1^{(m_1)}(G_1^{(m_1-1)}(...G_1^{(2)}(G_1^{(1)}(z)))) \\ G_2(z) = G_2^{(m_2)}(G_2^{(m_2-1)}(...G_2^{(2)}(G_2^{(1)}(z)))) \end{cases} \tag{3.6}$$

where $G_1^{(i)}$ and $G_2^{(i)}$ are the ith layers of G_1 and G_2 and m_1 and m_2 are the numbers of layers in G_1 and G_2. In our training process, m_1 and m_2 need not to be the same. In traditional discriminative deep neural network, the feature information is extracted from low-level features in first layers to the high-level features in last layers. While, through multi-layer perceptron operations, our two generator models decode the information with an opposite flow direction from abstract concepts to more material details.

Discriminative Models: Every generated sample has a corresponding class label and the discriminator gives both a probability distribution over dataset and a probability distribution over the class labels. We put both the original samples and the adversarial samples into D for the discriminator training, D output multiple output values between 0 and 1. In this process, if the training samples $\boldsymbol{x}$ is the positive/or real data, the discriminant D ensures the output value is similar with the trained corresponding value, which represents the input data is the positive/or real, while output values close to 0 indicates the input data is the negative/or fake. The discriminant D equals a classifier with supervision situation, which

returns to 1 or 0. Let D be the discriminative model given by:

$$D(\boldsymbol{x}) = D^{(n)}(D^{(n-1)}(...D^{(2)}(D^{(1)}(\boldsymbol{x})))) \tag{3.7}$$

where $D^{(i)}$ is the ith layer of D and n is the number of layers. The discriminator maps each input image to a probability score which indicates the input is drawn from the positive data or the negative data. In this process, the first layer of the discriminative model extracts low-level features, while the last layer extracts high-level features.

Learning: The Semi-Coupled-GANs framework corresponds to a constrained minimax game given by

$$\begin{aligned}\max_{D} \min_{G_1,G_2} V(G_1,G_2,D) = E_{x\sim p_{x_{data}}}[\log D(\boldsymbol{x} \mid \boldsymbol{y})] + E_{z\sim p_z}[\log(1-D(G_1(\boldsymbol{z})))] \\ + E_{z\sim p_z}[\log(1-D(G_2(\boldsymbol{z})))]\end{aligned} \tag{3.8}$$

There are two independent generators in Equation. 3.8, $E_{z\sim p_z}[\log(1-D(G_1(\boldsymbol{z})))]$ and $E_{z\sim p_z}[\log(1-D(G_2(\boldsymbol{z})))]$, both of them share a same discriminator $E_{x\sim p_{x_{data}}}[\log D(\boldsymbol{x} \mid \boldsymbol{y})]$. The two generative models synthesize a pair of adversarial samples for confusing the discriminative models. The discriminator gives both a probability distribution over image data and a probability distribution over the class labels, $D(\boldsymbol{x} \mid \boldsymbol{y})$. Here, there are four kinds of samples for training the discriminator: the positive and negative samples from original images and their corresponding adversarial samples computed by two generators. The inputs discriminative model is data and corresponding labels. Similar to GAN, our SCGANs can be trained by back propagation with the alternating gradient update steps.

3.3.1.3 Quality Estimation

For a given cardiac volume, a dissimilarity score is computed for each representative visual attribute - MAS and MBS. Any visual attributes with a score below an optimal threshold is classified as an artifact. After computing the visual attributes, we could verify the cardiac MRI quality based on the corresponding attributes scores. Let $x_{target} = P_{MAS}(\boldsymbol{X}_{target})$ and $y_{target} = P_{MBS}(\boldsymbol{X}_{target})$ be the outputs of the discriminator. If the quality of target cardiac volume $\boldsymbol{X}_{target}$ is good, the values $P_{MAS}(\boldsymbol{X}_{target})$ and $P_{MBS}(\boldsymbol{X}_{target})$ from the target cardiac volume should be similar with the trained corresponding positive attribute values. We combine the output values so the verification classifier Q can make sense of the data. To address the problem, we use the concatenation of these tuples for both MAS and MBS

attribute classifier outputs form the input to the verification classifier Q [120]. Finally, putting both terms together yields the tuples $q(S_{target})$:

$$q(S_{target}) = Q(< p_{MAS}, p_{MBS} >) \tag{3.9}$$

Training Q requires pairs of positive examples and negative examples. For the classification function, we use SVM with an RBF kernel for $\boldsymbol{X}$, trained using libsvm [27] with the default parameters of $C = 1$ and $\gamma = 1/ndims$, where $ndims$ is the dimensionality of $< p_{MAS}, p_{MBS} >$.

3.3.2 Experiment and Related Analysis

Data specifications: In the UK Biobank (UKBB) dataset, we have 3400 subjects, each with 50 time points covering the heart from the base to apex. We use the endocardial contour as the main characteristic to identify the apical, middle and basal slices. For example, we can find the Left Ventricular Outflow Tract (LVOT) in the basal slice. In other slices, LVOT is nonexistant. As for the apical slice, we define it as the LV cavity is still visible at end-systole. Besides the basal slice and apical slice, we can consider the rest slices as the middle slices. To obtain the negative samples, we choose the middle slice as the negative samples for each attribute learning.

Experimental set-up: All experiments used TensorFlow [1] on GPUs. With all 50 time points consideration for each subject, we can obtain 170,000 and regarded as the ground truth in our experiments. The architecture of the two generators G_1 and G_2 are consisted of several 'deconvolution' layers that transform the noise z and class c into an image [174]. We train the model architecture for generating images at 120×120 spatial resolutions. The discriminator D is a deep convolutional neural network with a Leaky ReLU nonlinearity [152]. In our experiment, 10-fold cross-validation method is used to evaluate the final performance of our attribute classifiers. To evaluate the classification algorithms, we use Accuracy, Precision Rate and Recall Rate defined as: *Accuracy = (TP+TN)/(TP+FP+TN+FN)*, *Precision Rate = TP/(TP + FP)* and *Recall Rate = TP/(TP + FN)*. Where TP, TN, FP, and FN are the numbers of the true positive, true negative, false positive and false negative samples, respectively.

Performance and Discussion: We evaluate the quality of our semi-supervi-sed representation learning algorithms by applying it as a feature extractor on supervised datasets. Table 1 shows the test performance on UK Biobank Dataset with the state-of-art deep learning methods. With supervised deep learning methods, 2D CNN, it achieved accuracies with 77.5% and 74.9%. With adversarial learning approach, traditional GAN, the

Table 3.2: The accuracy, precision rate and recall rate between the state-of-art deep learning approaches and our method.

Method	Accuracy		Precision Rate		Recall Rate	
	MAS	MBS	MAS	MBS	MAS	MBS
2D CNN	77.5±0.7%	74.9±0.6%	82.6±0.7%	74.9±0.8%	87.7±0.8%	87.8±0.9%
3D CNN	93.1±0.6%	91.8±0.7%	90.1±0.6%	87.3±0.7%	89.9±0.7%	93.3±0.8%
GAN	90.4±0.7%	88.1±0.6%	85.9±0.5%	88.5±0.6%	89.1±0.4%	90.6±0.6%
Our SC-GAN	92.5±0.5%	89.3±0.4%	87.6±0.4%	89.1±0.3%	90.5±0.5%	91.7±0.4%

results are much better with 90.4% and 88.1% accuracies. Compared with the above two methods, our SCGANs achieved performance with significant increase, 92.5% and 89.3% accuracies. This is despite the state of the art models having no ability to discriminate the adversarial samples, whereas our model requires to training the generative model to produce the adversarial examples and can correctly classify both unmodified and adversarial samples. It improves not only robustness to adversarial examples, but also generalization performance for original examples. Meanwhile, our SCGANs also achieved a comparable result with the 3D CNN, which indicates opportunity for future 3D image synthesis models.

Our attribute classifiers are trained using nine folds and then evaluated on the remaining fold, cycling through all ten folds. Receiver Operating Characteristic (ROC) curves are obtained by saving the classifier outputs for each test pair in all ten folds and then sliding a threshold over all output values to obtain different false positive/detection rates. In Figure 3.9, we demonstrate the ROC curve to show that our adversarial training (SCGANs) method can achieve ideal results. These results reinforce that adversarial examples are powerful samples for attribute leaning. In Figure 3.9 we can see our proposed method can correctly classify a few challenging samples (True Positive) and adversarial samples (False Negative). Experimental results obtained confirm that adversarial training approach makes the model more robust to adversarial examples and generalization performance for original examples. Although the results show that the accuracy of the proposed method is slightly lower but comparable to that of 3D CNN, our SCGAN can reduce the computation cost, which is especially important in population imaging. We provided the Figure 3.8 to show the training time of each method for 2,000 images (1,000 positive vs. 1,000 negative).

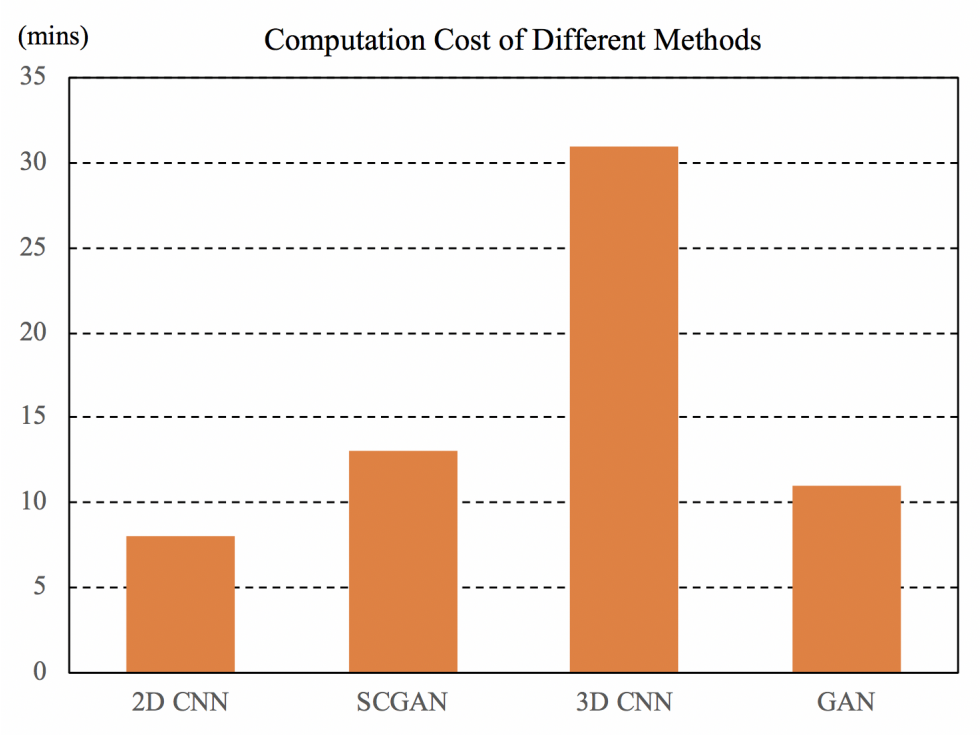


Figure 3.8: The computation cost comparison of different methods.

3.3.3 Conclusion

In this study, we tackled the problem of defining missing apical and basal slices in large imaging databases. We illustrated the concept by proposing a SCGANs to CMR image studies from the UK Biobank pilot datasets. By training the classifier with the adversarial examples, our model can achieve a significant improvement in attribute representation. A well-trained attribute classifiers are performed on the candidates to corresponding categories. We also validated our model by comparing with traditional deep learning methods and applying them to UK Biobank data sets. The proposed model shows a high consistency with human perception and becomes superior compared to the state-of-the-art methods, showing its high potential. Our proposed semi-couple-GANs can also be easily applied and boost the results for other detection and segmentation tasks in medical image analysis.

3.4 Limitations and Discussion

Multi-center studies are becoming more and more important (and technically feasible). On the one hand, this requires techniques to compensate for inter-site differences in data acquisition. On the other hand, pure quality assessment and discarding of data with insufficient quality is of high necessity as well, and computer-aided methods to achieve this

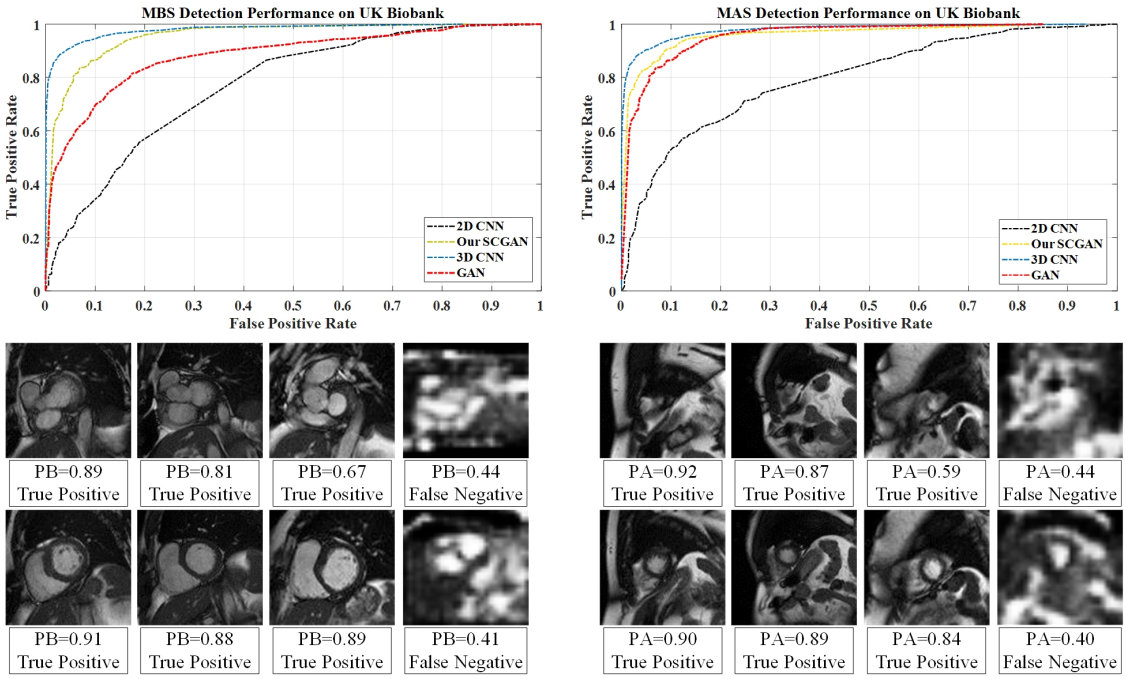


Figure 3.9: MAS and MBS detection performance (Top) and sample test slices and their probability values (Bottom). PA means the Probability value of being Apical slice; PB means the Probability value of being Basal slice.

are desirable if they operate with near-human or human-surpassing performance. Our first proposed approach used two fully independent CNN tracks for binary classification of apical and basal slices in the UKBB dataset, which has high relevance and addresses a very important problem. Our general approach of quality-aware image analysis is very original, few works exist on this important topic and could be very relevant also for other image modalities such as ultrasound. In terms of methodology, CNNs are relatively novel in the medical domain, a CNN approach is applied in different architectural configurations (3, 5 and 7 layers deep) and compared to SAEs and DBNs. Although we constructed a dataset (*i.e.*, including 100 annotated cardiac volumes), compared with the natural image domain which usually employs millions of training samples (e.g., ImageNet challenge provides 1.2 million images [118]), we still face the risk of over-fitting when training the 2D CNN models. Meanwhile, our current CNN based method did not use the adaptation to the 3D nature of the data, which could help in very difficult-to-assess borderline cases of MAS/MBS. A few neighboring slices could provide critical information. Especially in borderline cases (exactly those which are non-obvious and thus also difficult for humans), human observers would scroll through a few slices at the top/bottom to observe how anatomical boundaries are developing in 3D, before making a decision. In the future work, I would take sev-

eral slices instead, in order to give the CNN some 3D context. One potential approach to achieve this would be using parallel CNNs which are merged at the fully connected output layers.

Meanwhile, we proposed the second method called SCGANs using the GAN architecture to identify cardiac MR volumes which do not cover the heart in its entirety. A generative adversarial network was used with a pair of generators (one for negative examples, one for positive examples) followed by a shared discriminator. The proposed framework has two stages, First, the SCGANs generate adversarial examples and extract features from the CMR images subsequently, these image attributes are used to detect missing basal and apical slices. This method provide us a approach to generate more data for training the deep learning network and experimental results obtained confirm that adversarial training approach makes the model more robust to adversarial examples and generalization performance for original examples. However, we can also see from the experiment results that the accuracy of the proposed method is slightly lower but comparable to that of 3D CNN, which means the 3D contextual information is more important for the missing slice detection. Thus, constructing the 3D deep learning network is the key in our next step work.

In addition to construct a larger dataset and implement 3D neural network in the next step work, we also need to know what is the human error rate actually, and how long does the classification task take for a human? For a trained examiner, checking the ultimate two slices of a DICOM image stack should take only a few seconds, and should be very reliable, even under fatigue. Precision of 79.4% and recall of 88% of our proposed CNN method still seems like uncertainty, which does not yet provide any benefit for cross-center studies like the UK Biobank dataset. There are several ways to improve the approach and fix the problems we discussed in this section, and I will mention and give the solutions in the following chapter.

3.5 Conclusion

In this chapter, we have presented a novel cardiac image quality assessment problem, namely LV coverage assessment, that is helpful to calculate the cardiac parameters between different imaging modalities. The characters of LV are useful to identify the position, which the slice belongs to, since the LV in each slice shows a different shape and size. By testing the top/or bottom slices in cardiac MRI volumes, we are able to identify the missing slice categories. Recently, convolutional neural networks (CNNs) have been widely developed and used as a learning data-driven features approach for shape differences detection. In this

chapter, we proposed two methods for LV coverage assessment: (1) CNN based framework, and (2) Semi-coupled generative adversarial networks (SCGANs).

Recent work [110] has shown that the learned features extracted by CNN can be used for image quality assessment. Our initial experiments (first method) on the UKBB dataset have shown that CNNs are able to learn more distinctive features and suppress false positive detections, compared with other deep learning methods. However, the current CNN based structure and the SCGAN did not utilize the 3D information in cardiac volumes, which is important for discriminative feature learning. Thus, our follow-up work could investigate 3D CNN and the CNN based multi-modal feature space learning. This way one could potentially extract an optimal feature space to better identify the multi-modal correspondences. Although LV shape representation is experimentally shown to be a reasonable approach to solve this problem [148], there is no need to hand-craft the intermediate image representations, which can be learned implicitly through CNNs. In Chapter 4, we extend the proposed CNN representation by introducing a new 3D CNN model, namely Fisher Discriminative 3D CNN. The new approach enables us to learn population specific classification models, which results in improved accuracy with minimized within-class scatter and maximized between-class scatter.

Chapter 4

Automatic Assessment of Full Left Ventricular Coverage in Cardiac Cine Magnetic Resonance Imaging with Fisher-Discriminative 3D CNN

This chapter is based on:

- **Le Zhang**, Ali Gooya, Marco Pereañez, Bo Dong, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, Automatic Assessment of Full Left Ventricular Coverage in Cardiac Cine Magnetic Resonance Imaging with Fisher Discriminative 3D CNN, *IEEE Transactions on Biomedical Engineering*, 66(7), 1975-1986 (July 2019).

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, Image Quality Assessment for Population Cardiac MRI, *Deep Learning and Convolutional Neural Networks for Medical Image Computing.* Springer, Cham, 2019. (In Press)

Authors' contributions: L.Z., A.G., M.P., B.D. and A.F.F. conceived and designed the study; S.E.P., S.N. and S.P. provided support on clinical aspects and he also provided the UK Biobank data resource to be used for training and testing; L.Z. designed the method, performed data analysis and wrote the manuscript. All authors read and approved the manuscript.

Abstract: Cardiovascular magnetic resonance (CMR) imaging is a standard imaging modality for assessing cardiovascular diseases. Full coverage of the left ventricle (LV), from base to apex, is a basic criterion for CMR image quality and necessary for accurate measurement of cardiac volume and functional assessment. Incomplete coverage of the LV is identified through visual inspection, which is time-consuming and usually done retrospectively in the assessment of large imaging cohorts. This chapter proposes a novel automatic method for determining LV coverage from CMR images by using Fisher-discriminative three-dimensional (FD3D) convolutional neural networks (CNNs). In contrast to our previous method employing 2D CNNs, this approach utilizes spatial contextual information in CMR volumes, extracts more representative high-level features and enhances the discriminative capacity of the baseline 2D CNN learning framework, thus achieving superior detection accuracy. A two-stage framework is proposed to identify missing basal and apical slices in measurements of CMR volume. First, the FD3D CNN extracts high-level features from the CMR stacks. These image representations are then used to detect the missing basal and apical slices. Compared to the traditional 3D CNN strategy, the proposed FD3D CNN minimizes within-class scatter and maximizes between-class scatter. We performed extensive experiments to validate the proposed method on more than 5,000 independent volumetric CMR scans from the UK Biobank study, achieving low error rates for missing basal/apical slice detection (4.9%/4.6%). The proposed method can also be adopted for assessing LV coverage for other types of CMR image data.

Keywords: 3D convolutional neural network · LV coverage · image-quality assessment · population image analysis · Fisher discriminant criterion.

4.1 Introduction

Left ventricular (LV) cardiac anatomy and function are widely used in the field of cardiac medicine for diagnosis and monitoring disease progression and for assessing the patient's response to cardiac surgery and interventional procedures. Cardiac ultrasound (US) and cardiac magnetic resonance (CMR) imaging are arguably the most widespread techniques for diagnostic imaging of the heart. For population imaging studies, however, CMR remains the modality of choice. CMR is a single technique that provides access to cardiac anatomy and non-invasive measurements of cardiac function [182]. In large population imaging studies or assessment of patient cohorts from large clinical trials, the quantification of LV anatomy and function requires automatic image quality assessment and tools for image analysis. One basic criterion for cardiac image quality is LV coverage and detection

of missing apical and basal CMR slices [118]. CMR may display incomplete LV coverage because of insufficient radiographer experience in planning a scan, natural cardiac muscle contraction, breathing motion, and imperfect triggering, all of which pose challenges in efforts at quantitative LV characterisation and accurate diagnosis [188]. For example, missing basal slices affect calculations of LV volume and derived LV functional measures such as ejection fraction and cardiac output. Even if scout images are acquired, in order to centre the LV in view and minimize this issue, incomplete coverage may result at any point throughout the cardiac cycle because of changes in patient breathing and cardiac motion. Image quality assessment is traditionally performed by radiographers who ensure that patients do not leave the scanner without providing diagnostically interpretable data. However, there are limits to human attention. With CMR examinations becoming less expensive and increasingly commissioned, scanning loads at some centres may be insufficient to maintain consistent standards. Quality assessment is of particular importance in large-scale population imaging studies, where data are acquired across different imaging sites before core lab analysis. For example, large volumes of data may be stored without being checked by experienced staff prior to analysis [62] [253]. Automatic methods for these repetitive quality assurance tasks provide the required consistency and reliability.

To ensure consistent quantification of CMR data, automatic assessment of complete LV coverage is the first step. LV coverage is assessed by visual inspection of CMR image sequences, which is a subjective, repetitive, error-prone, and time-consuming process [9]. Automatic coverage assessment is required to promptly intervene and correct data acquisition, and/or discard images with incomplete LV coverage whose analysis would otherwise impair any statistics aggregated over the cohort. The most common causes of incomplete LV coverage are lack of a basal slice (no atrial chamber visible in end-systole, hence no certainty that the base of the heart is completely covered) and lack of an apical slice (LV cavity remains visible at end-systole). According to the criteria used in [118] for CMR quality assessment, a missing basal slice carries a higher penalty than a missing apical slice, given its impact on LV volume computation. Although technological developments in magnetic resonance imaging (MRI) hardware and pulse sequences have led to faster CMR acquisitions, challenges remain with regard to ensuring full heart coverage and motion compensation. In the UK Biobank's CMR protocol, for instance, incomplete heart coverage is the reason for flagging 4% of all CMR examinations as providing unreliable or non-analysable image data [21]. While 4% may seem to be a small proportion, the challenge is to automatically sift through the entire database to identify and exclude those cases from further quantitative analysis. Methods for the objective detection of basal and apical imaging planes are

relevant in this context, as their absence affects diagnostic accuracy as well as anatomical and functional LV quantification.

In the field of video processing, Automatic Image Quality Assessment (AIQA) is a well-developed corpus of techniques concerned with detecting image distortions characteristic of multimedia communications [202] [264]. These distortions generally differ from those affecting medical images. No-reference-based image quality assessment (NR-IQA) [87] [164] is relevant for medical imaging data. While there is relatively easy access to abundant data sets of mixed quality, it is not possible to collect data without some level of image degradation or artefacts. Practical CMR image-processing applications do not provide perfect versions of incomplete LV coverage images, but rather, only the image to be assessed. While assessments attempt to highlight differences in our assessed data set regarding a hypothetical high-quality image [110], the final image quality is estimated solely based on the characteristics of the assessed image.

UK Biobank's CMR acquisitions are performed on a clinical wide bore 1.5T scanner (MAGNETOM Aera, Syngo Platform VD13A; Siemens Healthcare, Erlangen, Germany) and include piloting, sagittal, transverse, and coronal partial coverage of the chest and abdomen. For measuring the cardiac function, three long-axis cines are acquired (viz. horizontal long-axis (HLA), vertical long-axis (VLA), and LVOT in both sagittal and coronal views). In addition, a complete SA stack is acquired. All acquisitions use balanced steady-state free precession (bSSFP) MRI sequences, attempting full coverage of the LV and right ventricle [183]. In this study, we will focus on SA bSSFP cine CMR data. To date, more than 18,800 volunteers have been scanned. Voxel and matrix size of these CMR images are, respectively, $1.8 \times 1.8 \times 8.0\text{mm}^3$, and 208×187 with, approximately, 10 slices per volume. Each volumetric sequence contains about 50 cardiac phases.

Quality-scored cardiac MRI data are available for approximately 5,000 volunteers of the UK Biobank (UKBB) imaging resource. Following visual inspection, manual annotation was carried out with a simple three-grade quality score [21]: (1) optimal quality for diagnosis, (2) suboptimal quality yet analysable and (3) bad quality and diagnostically unusable. In 5,065 SA cine CMR from the same number of volunteers, 4,361 sequences correspond to a quality score of 1, an additional 527 sequences have a quality score of 2, and the remaining 177 sequences have a quality score of 3. All datasets with optimal quality (score 1) had full coverage of the heart from base (LVOT existing) to apex (LV cavity still visible at end-systole). These data were used to construct the ground-truth classes for our experiments. Note that having full coverage should not be confused with having top/bottom slices corresponding exactly to the base/apex.

The current standard operating procedure in the UK Biobank, for instance, involves the detection of missing basal/apical slices based on visual assessment by experts. Few methods have been developed for automating this process, and prior work mostly adopted approaches that require segmenting short-axis slices of LV [16] [279] or landmark localization [90] [145] [49]. However, fast full LV coverage detection as the first step of an image quantification pipeline is largely unexplored. Hoffmann et al. pioneered this field [90] by initially localizing the heart in raw data prior to applying computer-aided diagnosis algorithms. Lu et al. [147] proposed an approach to locate LV and prescribe long/short-axis views before MR image acquisition, which could be used to evaluate cardiac coverage in short-axis views. These methods detect missing basal/apical slices and largely rely on the quality of LV segmentation and localization. de Vos et al. [48] proposed a method that automatically identifies a slice of interest (SOI) in 3D images. A ConvNet regressor was trained to determine the distance between each 2D slice and the SOI. However, this solution does not consider 3D contextual information contained across slices.

The characteristics of the LV are useful in identifying the position that the slice belongs to, since the LV in each slice shows a different shape and size. For example, the LV shape is approximately circular in mid-slices, while it is more elliptical in basal slices (Fig. 1.6). Recent work [67], [259] has focused on learning data-driven features to more accurately detect shape differences. Among them, 3D convolutional neural networks (CNNs) are one of the most regularly used deep-learning schemes to meet the challenges of discriminative shape detection [237] [136]. Roth et al. [196] and Prasoon et al. [187] adapted 2D CNNs for processing 3D volumetric data. However, these studies reported having difficulties when attempting to employ 3D CNN on their data, since they often lack sufficient training samples and computational resources to learn accurate 3D models. Although some authors [108] [241] have utilized 3D CNNs to process medical images, their architectural settings, convolution kernels, and prediction score volumes have not been disclosed in the detail required to reproduce their results [58]. Some exceptions, however, include the work of Kamnitsas et al. [109], who devised an effective dense training scheme based on 3D CNNs for brain lesion segmentation and dealing with the computational burden of processing 3D medical scans. Moreover, the 3D U-Net architecture of Cicek et al. [38] takes 3D volumes as input and produces volumetric image segmentation. The architecture and data augmentation of the U-net allow learning models with very good generalization performance from only a few annotated samples. Owing to the success of 3D deep neural networks in medical image segmentation, we are motivated to devise an end-to-end network optimization without requiring manual annotations of the visual image quality. Meanwhile, we seek

features maximally affected by partial image artefacts, which are also not very sensitive to variability related to the intrinsic anatomy or image modality at hand.

In this chapter, we focus on the analysis of short-axis (SA) cine MRI, although the technique can also be generalized to long-axis images. We aim to identify missing apical slices (MAS) and/or basal slices (MBS) in 3D cardiac MRI volumes. In our previous work, we used a 2D CNN constructed on single-slice images and processed them sequentially [271]. However, this solution ignores contextual information contained across slices providing inferior performance compared to a 3D analysis. We assume that 3D CNNs can easily and effectively deal with within-class variability and between-class similarity, which are important sources of the detection error [36]. We seek to learn a feature representation that achieves reliable classification results even with a small amount of training data or a small number of iterations. In this chapter, we address incomplete LV coverage detection using a Fisher-discriminative 3D (FD3D) CNN, which utilizes 3D convolution kernels and exploits the spatial contextual information in volumetric data. The proposed FD3D CNN uses the Fisher discriminant criterion [266] on the fully connected layer to render features more discriminative and insensitive to geometric structural variations.

To the best of our knowledge, this is the first study tackling the problem of automatic detection of missing basal and apical slices on a CMR dataset as extensive and challenging as the UK Biobank. Besides introducing a novel FD3D CNN architecture, we propose an effective cascaded detection strategy for incomplete coverage identification. In the first stage, we train two separate FD3D CNN classifiers to detect the absence of basal and apical slices. In the second stage, we combine the classification results from stage 1 to determine the type of incomplete coverage found on the image.

The rest of this chapter is organized as follows. Section 4.2 introduces the proposed FD3D CNN architecture and explains the learning strategy for its parameters. Section 4.3 presents experimental materials and metrics. Section 4.4 describes the experimental design and classification results. Further analysis and discussion of the proposed method are provided in Section 4.5. Conclusions are presented in Section 4.6.

4.2 Full LV Coverage Detection Method

4.2.1 Problem Formulation

During image acquisition, a sufficient margin ought to be left above and below the LV cavity according to the established guidelines [209]. However, some image volumes may lack sufficient information at the apical and basal levels, which can hamper or bias the

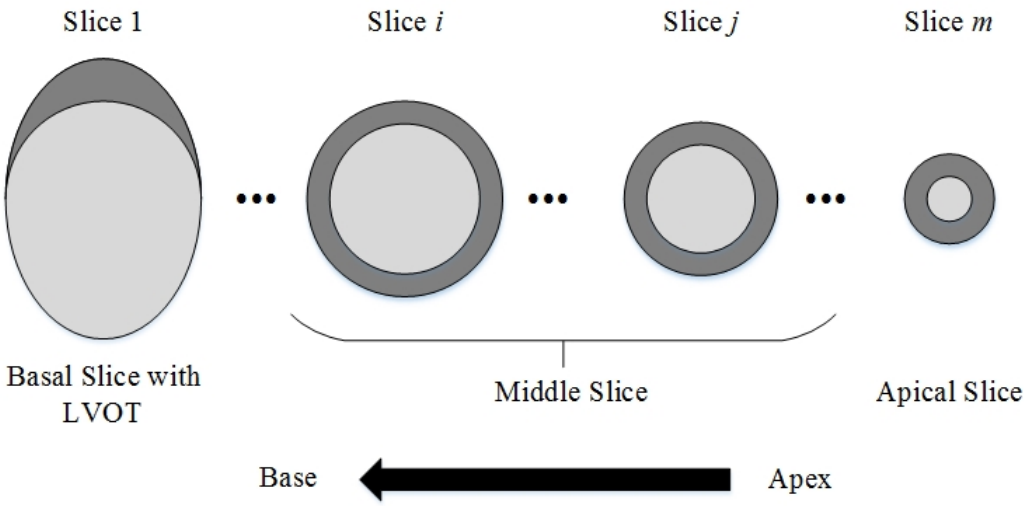


Figure 4.1: Schematic LV shapes showing blood pool (light gray) and myocardium (dark gray) for different slices from apex to base. Slice 1 (left) shows LVOT, which identifies the basal slice.

subsequent statistical analysis of cardiac structural and functional parameters in population imaging [184] [156]. In many LV quantification approaches, the LV cross section is approximated using simple quasi-circular models [37] [161]. These methods can produce a good approximation on LV mid-slices, but not on slices containing the left ventricular outflow tract (LVOT), which is at or near the basal slice. Therefore, in our approach, we treat the blood pool cross-section as a distinct model. Figure 4.1 depicts the LV shape of several slices in one cardiac volume from the apex to base. In volumes with missing basal slice, LVOT is usually not present.

We use a vector $\mathbf{s}$ to represent pixel values in each slice. A 3D cardiac MRI volume $\mathbf{V}$ with full coverage with n slices can be described as follows:

$$\mathbf{V} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n]. \tag{4.1}$$

Each cardiac volume, $\mathbf{V} = [\mathbf{s}_p, ..., \mathbf{s}_q], p \leq q \in [1, n]$, can have a different or same number of slices but cover a different portion of the LV.

To guarantee accurate cardiac volumetry and functional measurements [118], full LV coverage is a basic requirement [156]. To address this problem, we propose a two stage detection system that first computes image intensity representations by a FD3D CNN model and then detects missing slices based on these representations. In the first stage, we encode spatial contextual information and hierarchically extract high-level features, which indicate intensity representations. Our FD3D CNN model is equipped with a fully connected Fisher discriminative layer (F2) that takes the output of the fully connected layer (F1) as input. In

the second stage, independent detection of any missing basal and apical slices is performed and the results are combined to provide the final coverage assessment.

4.2.2 Three-dimensional Intensity Representations

Lu et al. [149] proposed a pattern recognition technique built on intra-segment correlation, using a normalization scheme, which maps each LV slice to polar coordinates with fixed size, shape level, and position. Intensity information and slice position are relevant even with incomplete LV coverage detection. In our chapter, we define intensity representation for the missing slice in a high-level feature space where slices of cardiac MRI are used to construct a representation of intensity. Each slice of the 3D volume is accounted for and the similarity of neighboring slices determines the difference of the 3D intensity distribution. Different characteristics in each slice and contextual information about spatial relation between slices are used to compute intensity representations.

Which 3D intensity representations? Our intensity representations are computed as a feature distribution matrix, which integrates information about LV shape and size. We detect incomplete LV coverage by image classification using the distribution matrix. We define two classes: missing apical slice (MAS) and missing basal slice (MBS).

Given a particular describable visual representation, we can formalize our notion of 3D intensity representations based on Equation 4.1. For example, if we are looking at the volume from base to apex, MAS and MBS can be formalized as follows:

$$\begin{cases} \mathbf{V}_{MBS} = [\mathbf{s}_q, ..., \mathbf{s}_n], \\ \mathbf{V}_{MAS} = [\mathbf{s}_1, ..., \mathbf{s}_p], \end{cases} \tag{4.2}$$

where, $p, q \in (1, n)$, $\mathbf{s}_1$ is the basal slice and $\mathbf{s}_n$ is the apical slice. Our intensity representations classifiers can be thought of functions $f(\cdot)$ for mapping 3D stacks $\mathbf{V}$ to real value p_i. A positive value of p_i indicates the presence or strength of the i^{th} representation, while negative values indicate its absence. Considering our intensity representations, if we define $\mathbf{V}_1$ and $\mathbf{V}_2$ as MBS and non-MBS samples, respectively, the representation function $f_{MBS}(\cdot)$ may map $\mathbf{V}_1$ to a positive value and $\mathbf{V}_2$ to a negative value. This is a binary classification function. Our 3D intensity representation classifiers are trained on the UK Biobank dataset as they provide reliable ground-truth labels based on visual inspection and manual annotation.

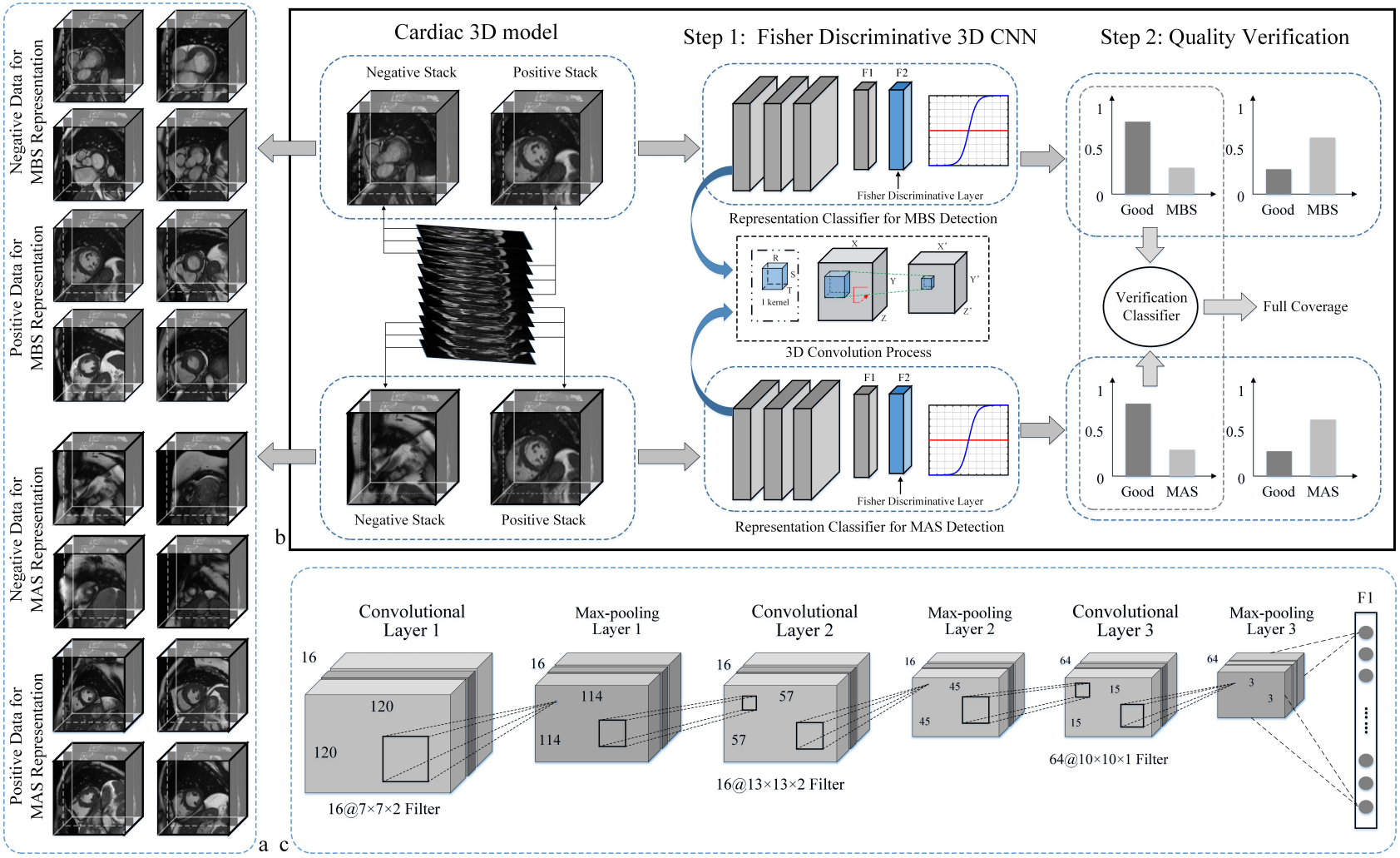


Figure 4.2: Whole assessment framework. *a:* Positive and negative training data for each representation classifier (MBS and MAS); *b:* Framework for our LV coverage assessment process; *c:* Structure and parameters of the 3D CNN used in panel b: Step 1.

4.2.3 Fisher Discriminative 3D CNN Model

In this subsection, we propose a FD3D CNN (shown in Figure 4.2b) to extract high-level features, which represent 3D intensity representations. Our FD3D CNN model is designed by adding a new Fisher-discriminative fully connected layer, F2, which uses the output of the previous layer, F1, as input. The new layer is then stacked onto a conventional 3D CNN. To maximize inter-class distances between learned features while minimizing intra-class distances of learned features, we train the newly added Fisher discriminative layer F2 on CNN features based on a Fisher discriminant criterion [266].

1) 3D CNN: Learning feature representations in three dimensions is important for later feature detection and image interpretation tasks in volumetric medical imagery. We employ 3D convolution kernels to encode richer spatial information in volumetric data. Here, feature maps are 3D blocks instead of 2D patches. Conventional 3D convolution is achieved by convolving a 3D kernel, with the cube formed by stacking multiple contiguous slices. With this construction, feature maps in the convolution layer are connected to multiple contiguous frames of the previous layer [103] [95]. Given an input $\mathbf{v}_k^l$, the 3D convolution layer output equates to a filtering operation with a filter $\mathbf{W}_{ik}^{l+1}$. Computation of the 3D feature volume $\mathbf{h}_i^{l+1}$ is given by:

$$\mathbf{h}_i^{l+1} = f\left(\sum_k \sum_{r=0}^{R-1} \sum_{s=0}^{S-1} \sum_{t=0}^{T-1} \mathbf{W}_{ik}^{l+1}(r,s,t)\mathbf{v}_k^l + b_k^{l+1}\right) \tag{4.3}$$

where $\mathbf{W}_{ik}^{l+1}(r,s,t)$ is the element-wise weight in the 3D convolution kernel, $\mathbf{W}_{ik}^{l+1}$ and b_k^{l+1} are the filter and bias terms connecting the feature maps of adjacent layers, and $f(\cdot)$ is the element-wise, non-linear activation function.

2) Fisher Discriminative 3D CNN: To boost the discriminative power of 3D CNN learned features, we impose a Fisher discrimination criterion [266] on them. Given the 3D input data $\mathbf{V}_i^t$, where i is the representation class, with $i = \{1,2\}$, corresponding to MAS and MBS; the superscript t in $\mathbf{V}_i^t$ indicates whether the representation is positive or negative, i.e., $t = \{0,1\}$; $\mathbf{V}_i^t = \left[\mathbf{v}_{i,1}^t, \mathbf{v}_{i,2}^t, ..., \mathbf{v}_{i,C}^t\right]$, $\mathbf{v}_{i,j}^t$ is the input data of j^{th} sample from class i, for $j = 1,2,...,C$. We denote $\mathbf{F}_{i,j}^t$ to be features in the fully-connected layer of the 3D CNN for class i and j^{th} sample. $\mathbf{F}$ is the extracted features of $\mathbf{V}$, which can be described as $\mathbf{F(V)}$. Using the Fisher criterion, discrimination is achieved by minimizing within-class scatter of $\mathbf{F}^t$, denoted by $S_w(\mathbf{F}^t)$, and maximizing between-class scatter of $\mathbf{F}^t$, denoted by $S_b(\mathbf{F}^t)$. $S_w(\mathbf{F}^t)$ and $S_b(\mathbf{F}^t)$ are defined as follows:

$$S_w(\mathbf{F}^t) = \sum_{i=1}^{I} \sum_{\mathbf{F}_{i,j}^t \in t} (\mathbf{F}_{i,j}^t - \mathbf{m}_i^t)(\mathbf{F}_{i,j}^t - \mathbf{m}_i^t)^T, \tag{4.4}$$

$$S_b(\mathbf{F}^t) = \sum_{i=1}^{I} n_i(\mathbf{m}_i^t - \mathbf{m}^t)(\mathbf{m}_i^t - \mathbf{m}^t)^T, \tag{4.5}$$

where $\mathbf{m}_i^t$ and $\mathbf{m}^t$ are mean vectors of $\mathbf{F}_i^t$ and $\mathbf{F}^t$, respectively, and n_i is the number of samples from class i. The Fisher discriminant regularization term $\Phi(\mathbf{F}^t)$ is defined as $\mathrm{tr}(S_w(\mathbf{F}^t)) - \mathrm{tr}(S_b(\mathbf{F}^t))$. To obtain a discriminative classification result with deep learned features, we propose modifying the objective function of the FD3D CNN model by inserting a Fisher discriminant regularization term:

$$\begin{aligned}\mathbf{J}^*(\mathbf{W},\mathbf{b}) = \arg\min_{\mathbf{W},\mathbf{b}} \frac{1}{I}\sum_{i=1}^{I} y^t \log a(\mathbf{V}_{i,j}^t;\mathbf{W},\mathbf{b}) + (1-y^t)\log\big(1-a(\mathbf{V}_{i,j}^t;\mathbf{W},\mathbf{b})\big) \\ + \frac{1}{2}\lambda\,\|\mathbf{W}\|_2^2 + \frac{1}{2}\eta\,(tr(S_w(\mathbf{F}^t)) - tr(S_b(\mathbf{F}^t))),\end{aligned} \tag{4.6}$$

where $\mathbf{J}^*$ is our new cost function that can minimize within-class scatter and maximize between-class scatter, and y is the output label. Output activation $a(\mathbf{V}_{i,j}^t;\mathbf{W},\mathbf{b}) = 1/(1+e^{-\mathbf{W}\mathbf{V}_{i,j}^t-\mathbf{b}})$ is typically restricted to the open interval $(0,1)$ by using a logistic sigmoid, which is parametrized by $\mathbf{W}$ and $\mathbf{b}$ on the j^{th} training sample. $\|\mathbf{W}\|_2^2$ is a penalty term to the loss function that prevents weights from getting too large and helps to prevent over-fitting. Weights in each layer can be adjusted toward target classes and utilize input data close to the corresponding classes in case of no large dataset or a small number of iteration. Here, $\lambda, \eta \in [0,1]$ are two trade-off parameters that control the relative importance of each term and are usually chosen by experiments, which can differ depending on different databases and network structures.

For intensity representation $\mathbf{V}_{i,j}^t$, we define:

$$\mathbf{J}(\mathbf{W},\mathbf{b}) = y^t \log a(\mathbf{V}_{i,j}^t;\mathbf{W},\mathbf{b}) + (1-y^t)\log\big(1-a(\mathbf{V}_{i,j}^t;\mathbf{W},\mathbf{b})\big), \tag{4.7}$$

$$\Phi(\mathbf{F}^t) = \frac{1}{2}\mathrm{tr}\big((\mathbf{F}_{i,j}^t - \mathbf{m}_i^t)(\mathbf{F}_{i,j}^t - \mathbf{m}_i^t)^T\big) - \frac{1}{2}\mathrm{tr}\big((\mathbf{m}_i^t - \mathbf{m}^t)(\mathbf{m}_i^t - \mathbf{m}^t)^T\big). \tag{4.8}$$

Once the new cost function is obtained, we can employ the gradient descent method [118] to solve this optimization problem. Our key problem is to calculate the error of output units, which consists of output errors from two sub-functions $\mathbf{J}(\mathbf{W},\mathbf{b})$ and $\Phi(\mathbf{F}^t)$. To update parameters $\mathbf{W}^t$ and $\mathbf{b}^t$, we first calculate the error $\delta_i^{L,t}$ (L is the output layer) of the output layer with forward propagation, and then adopt the back-propagation method [100] to calculate the error $\delta_i^{l,t}(l<L)$ for other layers. Partial derivatives of the overall cost function $\mathbf{J}^*(\mathbf{W},\mathbf{b})$ regarding $\mathbf{W}^t$ and $\mathbf{b}^t$ are:

$$\frac{\partial \mathbf{J}^*(\mathbf{W},\mathbf{b})}{\partial W^{l,t}} = \sum_{t=0}^{C}\sum_{F^t\in t}\frac{\partial \mathbf{J}(\mathbf{W}^t,\mathbf{b}^t)}{\partial W^{l,t}} + \eta\sum_{t=0}^{C}\sum_{F^t\in t}\frac{\partial \Phi(\mathbf{F}^t)}{\partial W}, \tag{4.9}$$

$$\frac{\partial \mathbf{J}^*(\mathbf{W},\mathbf{b})}{\partial b^{l,t}} = \sum_{t=0}^{C} \sum_{F^t \in t} \frac{\partial \mathbf{J}(\mathbf{W}^t,\mathbf{b}^t)}{\partial b^{l,t}} + \eta \sum_{t=0}^{C} \sum_{F^t \in t} \frac{\partial \Phi(\mathbf{F}^t)}{\partial b}. \tag{4.10}$$

$\mathbf{F}_l(\mathbf{V}^t_{i,j}) = \kappa(\mathbf{W} \cdot \mathbf{F}_{l-1}(\mathbf{V}^t_{i,j}) + \mathbf{b})$, where $\kappa(x) = max(0,x)$. As shown on the partial derivatives on equations (4.9) and (4.10), the parameters $(\mathbf{W},\mathbf{b})$ (weights and biases) of the Fisher regularization term $\Phi(\mathbf{F}_t)$ iteratively optimized during the network training process. In this stage, we use the 3D CNN model with architecture in Table 4.1. Algorithm 1 provides the pseudo-code to train this new network. In our 3D CNN implementation, a rectifier linear unit (ReLU) [123] is utilized as a non-linear activation function in layers C and F1.

Algorithm 1: FD3D CNN Training.

Input: input-target pairs ($\mathbf{v}^t_{i,j}$, $\mathbf{y}^t$), corresponding j^{th} pairs from class i, t indicates positive or negative sample; η.

Output: FD3D CNN weight and biases, respectively, $\mathbf{W} = [\mathbf{W}^{1,t}, \mathbf{W}^{2,t}, ..., \mathbf{W}^{l,t}]$ and $\mathbf{b} = [\mathbf{b}^{1,t}, \mathbf{b}^{2,t}, ..., \mathbf{b}^{l,t}]$.

Begin

Initialize $\mathbf{W}^t_{i,j}$ and $\mathbf{b}^t_{i,j}$

while *stopping criterion has not been met* **do**

1) Classification error:
$\text{argmin}_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{\mathrm{I}} y^t \log a(\mathbf{V}^t_{i,j};\mathbf{W},\mathbf{b}) + (1-y^t) \log\Big(1 - a(\mathbf{V}^t_{i,j};\mathbf{W},\mathbf{b})\Big)$.

2) Fisher discriminant: $\Phi(\mathbf{F}^t) = \text{tr}(S_w(\mathbf{F}^t)) - \text{tr}(S_b(\mathbf{F}^t))$.

3) Discriminative objective function: $\text{argmin}_{\mathbf{W},\mathbf{b}} \sum_{i=1}^{\mathrm{I}} y^t \log a(\mathbf{V}^t_{i,j};\mathbf{W},\mathbf{b}) + (1 - y^t) \log\Big(1 - a(\mathbf{V}^t_{i,j};\mathbf{W},\mathbf{b})\Big) + \frac{1}{2}\lambda \|\mathbf{W}^t\|_2^2 + \frac{1}{2}\eta\,\Phi$.

4) Update $\mathbf{W}^t_{i,j}$ and $\mathbf{b}^t_{i,j}$ with Equations (4.9) and (4.10).

end

return $\mathbf{W}^t_{i,j}$ and $\mathbf{b}^t_{i,j}$ until values of $\mathbf{J}^*(\mathbf{W},\mathbf{b})$ in successive iterations are close enough or the maximum number of iterations is reached.

End begin

4.3 Materials and Metrics

4.3.1 CMR Acquisition Protocol and Annotation

4.3.2 Training and Testing Set Definitions

Training set: To create a training dataset for learning intensity representations, we extract the three topmost slices as negative samples for MBS detection (i.e. containing the cardiac base), and the three bottom most slices as negative samples for MAS detection. To

Table 4.1: Architecture of the 3D Discriminative CNN Model

Layer	Kernel Size	Stride	Output size	Feature volumes
Input	–	–	$120 \times 120 \times 3$	1
C1	$7 \times 7 \times 2$	1	$114 \times 114 \times 2$	16
M1	$2 \times 2 \times 1$	2	$57 \times 57 \times 2$	16
C2	$13 \times 13 \times 2$	1	$45 \times 45 \times 1$	16
M2	$3 \times 3 \times 1$	1	$15 \times 15 \times 1$	16
C3	$10 \times 10 \times 1$	1	$6 \times 6 \times 1$	64
M3	$2 \times 2 \times 1$	1	$3 \times 3 \times 1$	64
F1	–	1	$1 \times 1 \times 1$	256
F2	–	1	$1 \times 1 \times 1$	4

Note: F2 is the Fisher Discriminant Layer.

create positive samples (i.e. not containing the cardiac base/apex), we choose three-slice blocks, each starting from the middle slice towards the base/apex for MBS/MAS detection training. We create the training set from images with optimal quality and with exclusively full coverage.

We train using three-slice stacks (or triplets) to model the 3D context. the average number of slices per image volume is approximately 10. During training, we extract four triplets (two samples including base/apex and two samples excluding the base/apex). To maximize inter-class separation, it is wise to avoid intersection between the training samples; for example, if we use four-slice stacks (for a ten-slice volume), there will be a two-slice overlap between basal positive/negative examples and the apical region. By choosing the proposed slice triplets, we ensure that there is no overlap and increase the discriminative power of the FD3D CNN. Another important observation that supports the choice of slice triplets is that the CMR scan volume is not acquired immediately. Instead, each slice is collected over several cardiac cycles leading to some degree of slice-to-slice misalignment. This effect is minimized when considering only slice triplets in contrast to using the full 3D volume.

Testing set: During testing, we extract every set of three adjacent slices from top to bottom for each volume and apply these triplets to intensity representation classifiers. Data with known MBS/MAS are created by manually removing the three topmost/bottom most slices from images with optimal quality, as in the training set.

During training and testing, three-slice stacks are input to the proposed FD3D CNN. Scores of the output layer can be interpreted as the probability that triplets correspond to negative or positive MBS/MAS. The final output is the combination of two CNN outputs (MBS and MAS). The three slice stacks input into our network are cropped centered images

of dimensions $120 \times 120 \times 3$ to extract the region of interest. Parameter setting of block-size determination is explained in Section 4.4.1.

4.3.3 Training Set Augmentation

To prevent over-fitting due to insufficient training data and to improve the detection rate of our algorithm, we employ data augmentation techniques to artificially enlarge our dataset [92] [165]. In our application, we augment the data by applying a discrete set of in-plane rotations and isotropic scalings to the training images. Unlike data augmentation choices made for natural image datasets where variability in location and pose of objects are relatively high, our data are comparatively constrained due to standard imaging protocols and gross patient positioning on the MRI scanner. We therefore chose a set of realistic rotations and scaling factors for MRI. Based on analysis of the in-plane orientation angle distribution for 5,000 subjects for which manual segmentations were available (and therefore LVRV angle can be computed), we found that LVRV orientation ranges between -45° and 45°. The set of rotations chosen was accordingly -45° and 45°, with two scaling factors of 0.75 and 1.25. This increases the number of training samples by a factor of four, while not adding significantly to the convergence time.

After data augmentation, we constructed 845,000 3D stacks comprised of 2D CMR slices from 3,380 sequences each with 50 cardiac phases, with a quality score of 1. These data are used for experiments in Section 4.4.1, 4.4.2, and 4.4.3. We set aside 981 sequences and data with quality scores of 2 and 3 for later use, as described Section 4.4.4. In our experiments, 10-fold cross-validation [120] was used to evaluate the performance of our system. To the best of our knowledge, this is the largest annotated dataset available to date for automatic CMR quality assessment.

4.3.4 Learning Performance Metrics

To evaluate the learning process, we use the following established classification metrics:

$$\text{Precision} = TP/(TP+FP), \tag{4.11}$$

$$\text{Sensitivity} = TP/(TP+FN), \tag{4.12}$$

$$\text{Error Rate} = (FP+FN)/N, \tag{4.13}$$

where TP, FP and FN are numbers of true-positive, false-positive and false-negative samples, respectively, and N represents the number of subjects in the test set.

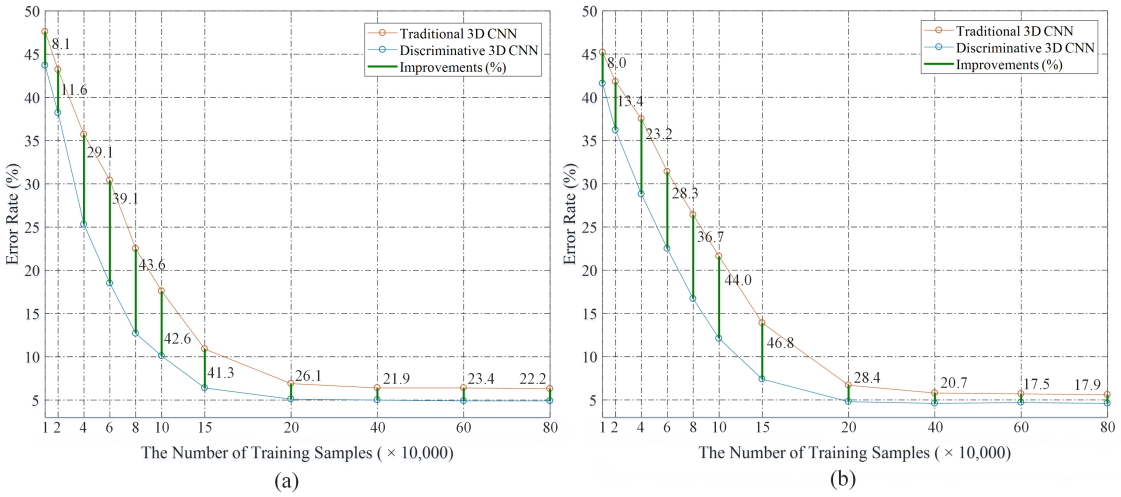


Figure 4.3: Error rates and improvements for increasingly larger training sets: (a) MBS detection, (b) MAS detection.

4.4 Experiments and Results

4.4.1 Performance Analysis

We experiment to characterize the performance of our FD3D CNN learning framework. The error (cost) functions used in learning (Equations 4.6 and 4.7) remain within this range $[0,1]$. In all experiments, the learning process was terminated when standard deviation of the error function over the last five iterations is smaller than $\sigma = 0.01$.

1) Hyper-parameter selection: LeCun *et al.* [210] and Salah *et al.* [203] used CNN to recognize handwritten digital numbers with different numbers of training samples on the MNIST dataset. Their results illustrated that, when reducing training samples, the recognition rate of the algorithm drops sharply. To demonstrate the behaviour of our FD3D CNN, we experiment with different percentages of training samples. We use *improvement* defined as $(1 - ER_{\mathrm{D}}/ER_{\mathrm{T}}) \times 100$ to benchmark our method against a traditional 3D CNN, where ER_{D} and ER_{T} are error rates of our FD3D CNN and the traditional 3D CNN, respectively. Error rates of MBS/MAS representation learning are shown in Figure 4.3, where our proposed method appears to achieve comparable results with less training data compared to the conventional 3D CNN. We choose 80% of the 845,000 as the training samples and perform testing on the remaining 20%. The results are shown in Table 4.2. Even when trained with fewer iterations, our method achieves better results than the traditional 3D CNN.

With sufficient training samples and iterations, most machine learning methods can improve their accuracy at a higher computational cost. However, we usually want to obtain a trained network as quickly as possible. This is especially important in population imaging

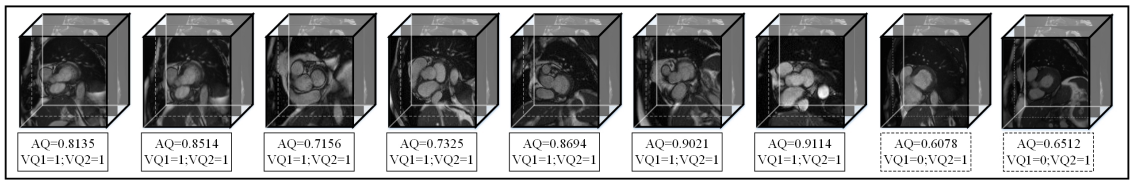
Table 4.2: Error rates versus Learning epochs

Epochs	Error Rate (%) Traditional 3D CNN (MBS/MAS)	Discriminative 3D CNN (MBS/MAS)	Improvement (%) (MBS/MAS)
1	32.4/30.7	28.8/27.4	11.1/10.8
10	25.4/24.2	19.2/17.6	24.4/27.3
20	19.2/18.7	11.3/10.8	41.1/42.2
30	12.7/13.1	8.3/8.6	34.6/34.4
40	6.3/5.6	4.9/4.6	22.2/17.9

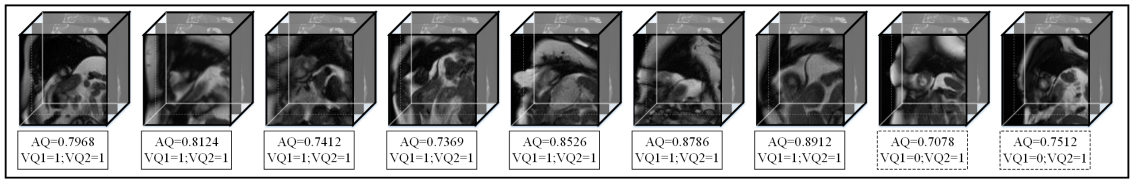
as new datasets can become available and retraining might be required. Rapid training is also a desirable feature during algorithmic development since finding an optimal architecture may require multiple training procedures for different parameter settings. We illustrate that our FD3D CNN has better error-reducting performance as a function of the number of training samples and iterations than other competing techniques.

A 3D CNN requires a suitable receptive field (i.e. input size) to achieve the best discrimination. Based on a random sample of 200 image volumes, we determine the smallest crop size that ensures the coverage of the LV structure compared to three block-size configurations, namely, $120 \times 120 \times 3$ (which removes redundant background information based on the central point of original images), $180 \times 180 \times 3$ (which is the original size as extracted and resized from the UK Biobank), and $80 \times 80 \times 3$, which mostly contains the LV at the centre. We test sizes smaller than the original block size of the classification model because we want to determine whether a larger input block with more contextual information can enhance the model's discriminative capacity. The results obtained with these settings are shown in Table 4.3. With a block size of $80 \times 80 \times 3$, MBS/MAS detection precision rate reaches 89.01% and 88.36%, respectively. The detection performance improves to a precision rate of 91.81% and 90.73% under block size $120 \times 120 \times 3$, demonstrating that increasing contextual information can enhance the discriminative capacity of 3D CNN. Without cropping, the detection precision rate decreases to 90.12% and 89.78% for MBS and MAS detection, respectively. This may have been because too much redundant contextual information clutters the actual LV signature, and hence degrades detection performance. Based on these experiments, we set block size to $120 \times 120 \times 3$, to achieve optimal detection performance.

Typical classification results using the proposed FD3D CNN architecture are shown in Figure 4.4. A few basal stacks (top row) and apical stacks (bottom row) in the test datasets



(a) Sample volumes for MBS testing with automatic quality (AQ), expert cardiologist (VQ1) and cardiac image expert's visual (VQ2) qualities.



(b) Sample volumes for MAS testing with automatic quality (AQ), expert cardiologist (VQ1) and cardiac image expert's visual (VQ2) qualities.

Figure 4.4: Sample test volumes and their AQ, expert cardiologist (VQ1) and cardiac image expert's visual (VQ2) qualities for MBS detection (top row) or MAS detection (bottom row) are shown. The left seven samples in each row show consistency between AQ and VQ1, which means our algorithm yields an accurate prediction; The right two samples in each row show the wrong quality prediction and show inconsistency between VQ1 and VQ2.

Table 4.3: Performance versus Block Size with Fisher Discrimination Criterion

Block Size	Precision		Sensitivity	
	MAS	MBS	MAS	MBS
$80 \times 80 \times 3$	89.01%	88.36%	88.24%	87.94%
$120 \times 120 \times 3$	**91.81%**	**90.73%**	**90.92%**	**90.25%**
$180 \times 180 \times 3$	90.12%	89.78%	89.63%	88.92%

with their AQ or corresponding posterior probability values are shown. High score values on the stack correspond to the likelihood of being a correct basal or apical triplet. Basal slices with existing LVOT indicate higher probability values of being correctly classified. This shows that the training captures the LVOT as a prominent feature in correctly positioned basal slices.

2) Comparison to other machine learning methods: We compare our framework with a traditional 3D CNN and with our previous 2D CNN study [271]. Table 4.4 lists the results for these architectures.The architecture of a traditional 3D CNN is similar to that of our FD3D CNN, replacing the Fisher layer (F2) with a traditional fully connected layer including 256 ReLU activation neurons. We use the same training and testing approaches for the 3D CNN and list the results obtained using the hand crafted features used in [148].

Table 4.4: Performance comparison of different learning models with learned and hand-crafted visual representations.

Method	Features	Precision (%)			Sensitivity (%)		
		MAS	MBS	$\overline{\text{MBS} \vee \text{MAS}}$	MAS	MBS	$\overline{\text{MBS} \vee \text{MAS}}$
FD3D CNN	Learned	**91.81 ±0.21**	**90.73 ±0.28**	**91.12 ±0.24**	**90.92 ±0.26**	**90.25 ±0.28**	**90.15 ±0.22**
3D CNN		89.12 ±0.36	89.32 ±0.34	89.20 ±0.31	89.42 ±0.31	89.47 ±0.30	89.25 ±0.29
2D CNN		81.61 ±0.56	74.10 ±0.58	79.42 ±0.62	88.73 ±0.49	88.75 ±0.51	88.01 ±0.56
Lu *et al*. [148]	Hand-crafted	37.60 ±1.22	45.68 ±1.36	56.92 ±1.71	67.43 ±0.92	74.56 ±1.32	63.25 ±1.79

In [148], the basal slice was identified following these steps: 1. Choose the mid-slice image as the start image and process each image sequentially in the basal direction. 2. Apply the optimal threshold method to convert the ROI to a binary image. 3. Identify the binary object with blood pool, which shows an elliptical shape. 4. Calculate the length of the major axis L of the ellipse that has the same normalized second central moments as the binary object. 5. If the ratio of the current to preceding L exceeds a predefined threshold (e.g. > 1.2 in this work), then a basal slice is identified; otherwise, the basal slice is missing. We use a similar method to identify the apical slice. We process each image sequentially from base to apex. If the ratio of the current to preceding L is smaller than a predefined threshold (e.g. < 0.2 in this study), an apical slice is detected; otherwise, the apical slice is missing. We employ this feature extraction procedure for prediction. The proposed FD3D CNN shows the best precision and sensitivity figures in each representation classifier, and full LV coverage detection performance.

4.4.2 Inter-Observer Reliability

To contextualize the results of automatic full LV coverage assessment, we compare it to the inter-observer full LV coverage detection rate obtained by expert readers. The inter-observer agreement [69] of human experts is evaluated by reassessing a subset of 200 random CMR datasets. The quality distribution levels in this randomly selected subset are compared to original data using Pearson's χ^2 goodness-of-fit test to confirm that it represents the original data distribution ($p > 0.05$). The reassessed samples demonstrate strong agreement with original qualities (Cohen's $\kappa = 0.76$, $p < 0.05$).

To show how our results can be compared to the expected human detection error rates, we present the error rates between an expert cardiologist (VQ1) and another cardiac image expert (VQ2) for 200 re assessed samples. The confusion matrix of VQ1 versus VQ2 is presented in Table 4.5. Use of the confusion matrix reveals 7 among the 200 re assessed samples with inconsistent quality assessment between VQ1 and VQ2. These findings show that the expert cardiologist's visual results conflict with the cardiac image expert's visual assessment only 3% of the time. As shown in Table 4.2 *epoch = 40*, our automatic algorithm's error rate is just below 5%, which shows excellent agreement with human expert assessments (two percentage points). Some examples of MBS/MAS test images are shown in Figure 4.4 (panels a and b correspondingly). We have intentionally chosen to show seven inter-observer agreement examples, plus two disagreement examples on each panel.

Table 4.5: Confusion matrix of the expert cardiologist (VQ1) and cardiac image expert's visual (VQ2) results. Grey numbers indicate number and ratio of correct estimates.

		VQ2			Correct
		MBS	MAS	$\overline{\text{MBS} \vee \text{MAS}}$	
VQ1	MBS	67	0	3	0.96
	MAS	0	65	2	0.97
	$\overline{\text{MBS} \vee \text{MAS}}$	1	1	61	0.97

4.4.3 Cross-database Performance: Sunnybrook Cardiac Dataset

We evaluate the generalization of the performance of our full LV coverage detection system on an independent database. We assess the sensitivity of our system to moderate changes in imaging conditions, scanner vendors, image resolution, etc. To this effect, we use Data Science Bowl Cardiac Challenge Data (Kaggle or Sunnybrook Cardiac dataset) [160]. This dataset comprises 1,120 cardiac MRI volumes. Cine steady state free precession (SSFP) MR short-axis (SAX) images are obtained with a 1.5T GE Signa CV/i MRI System (General Electric, Milwaukee, WI). All images are obtained during 10-15 second breath-holds with a temporal resolution of 20 cardiac phases over the heart cycle (scanned from the ED phase). Six to twelve SAX images are obtained from the atrioventricular ring to the apex (resolution $1.25 \times 1.25 \times 8\text{mm}^3$, thickness = 8mm). Gold-standard full LV coverage is obtained by an experienced reader and checked visually by inspecting slices from base to apex. Original volumes are used for full LV coverage detection and triplets of top and bottom slices are used, respectively, as negative examples for MBS and MAS. Positive examples of MBS/MAS are obtained from triples of mid-slices. This dataset is used as a test set for the FD3D CNN that was pre-trained with 800,000 volumes from the UK Biobank. Values for error, precision and sensitivity under various conditions are shown in Table 4.6.

Table 4.6: Cross-dataset performance: Kaggle dataset.

	Error (%)	Precision (%)	Sensitivity (%)
MAS	6.43	86.51	88.74
MBS	7.02	84.03	85.69
$\overline{\text{MBS} \vee \text{MAS}}$	6.64	85.74	87.01

4.4.4 Missing Slice Rate per Visual Quality Score

To gain insight into the relation between missing slice rates and visual quality scores achieved by experts [21], a third experiment is conducted. The system is trained on 3,380 random volumes from a total of 5,065. The testing set, as earlier indicated, has 1,685 CMR volumes distributed among the quality scores (from 1 to 3: 981, 527 and 177). Table 4.7 gives the percentages of the full LV coverage class for each quality score. CMR data with a quality score of 3 highly correlates with MBS, as missing basal slices highly affect accurate quantitative analysis in CMR.

Table 4.7: Missing Slice Rate per Visual Quality Score.

Quality Score	MAS (%)	MBS (%)	$\overline{\text{MBS} \vee \text{MAS}}$ (%)
1	1.7	0.6	97.7
2	74.7	24.0	1.3
3	18.0	80.4	1.6

4.4.5 Clinical Impact

To assess the impact of incomplete LV coverage in real applications, such as measurement of cardiac function based on blood volumes, we design an experiment where incomplete coverage is simulated and volume differences between full and incomplete coverage are measured. We also compute two commonly used indexes of the cardiac function derived from such volumes viz. stroke volume (SV) and ejection fraction (EF), and similarly report the differences between the full and incomplete coverages. For this experiment, we take 4,737 subjects for which manual annotations are available (both cardiac phase labels and full coverage labels), and systematically remove the basal and apical slices to generate incomplete MBS and MAS volumes. Then, we compute blood pool volumes at the ED and ES phases, and from these, we obtain SV and EF. Finally, the average volumes and indexes are computed across the sample, comparing full coverage and MBS/MAS. Table 4.8 shows that the largest effect of incomplete coverage is caused by MBS, where the missing slice reduces ED and ES volumes by an average of 12% and 20%, respectively. In turn, these differences cause a decrease in the computed SV by 6.7% and an increase in the EF by 3.9%. The absence of the apical slice has a smaller yet non-negligible impact on the volumes and derived indexes.

Table 4.8: Effect of incomplete cardiac coverage (MBS/MAS) on the End-diastolic, End-systolic, stroke volumes and ejection fraction. Values are shown as Mean ± standard deviations.

	Full	MBS	Effect(%)	MAS	Effect(%)
LVEDV(ml)	155.8±35.6	136.1±33.4	-12.6%	151.5±35.1	-2.7%
LVESV(ml)	66.8±21.2	53.0±19.0	-20.0%	64.3±20.9	-3.7%
LVSV(ml)	89.1±19.8	83.1±19.7	-6.7%	87.1±19.6	-2.2%
LVEF(%)	57.1±0.06	61.0±0.06	+3.9%	57.5±0.06	+0.4%

4.4.6 Implementation Considerations

The experiments reported here are conducted using the ConvNet library [51] on an Intel Xeon E5-1620 v3 @3.50GHz machine running Windows 10 with 32GB RAM and Nvidia Quadro K620 GPU. The networks are optimized using the gradient descent method [118] with the fllowing hyper-parameters: learning rate = 0.01, momentum = 0.9, drop-out rate = 0.1. Trainable weights are randomly initialized from a Gaussian distribution ($\mu = 0$, $\sigma = 0.01$) and updated with standard back-propagation. Models converge in about 6 hours when training is performed with 800,000 volumes with size 120×120×3. Testing is rapid and can process each volume in 3 seconds.

4.5 Discussion

Automatic identification of CMR volumes with incomplete LV coverage is important in high-throughput image analysis of population imaging. The acquisition of thousands of suboptimal CMR images for later image analysis can be avoided if such quality assessment is performed online and a system provides immediate feedback to technical staff when new images are acquired. Incomplete LV coverage influences the accuracy of anatomical and functional LV parameters of clinical interest. Manual annotation of LV coverage is laborious, time-consuming and error prone in current clinical routines. To automate this labour-intensive task, we propose an efficient and robust two-stage framework for the automatic detection of missing slices at the LV base and apex. In the first stage, we train a FD3D CNN that computes the corresponding intensity representation with high accuracy. It can qualify CMR volumes based on two representations, and can assist radiologists by automatically labelling the potentially incomplete volumes to mark them for closer inspection. The second stage robustly discriminates two quality categories (MBS and MAS),

based only on the intensity representation classifiers, which are then used to recognize new cardiac volumes with no further training. Specifically, to use the spatial information in volumetric data, we use 3D CNN with shared 3D convolution kernels. Meanwhile, a Fisher discriminant layer leads to small within-class scatter and large between-class scatter of feature vectors in that layer. Extensive experimental results illustrate the effectiveness and efficiency of our method: its performance is superior to that of other methods with obvious advantages.

In any AIQA system for population imaging, accuracy and robustness are key design criteria. These methods must work without many false positives or false negatives, and must cope with considerable variation in image quality. Most machine learning methods can improve their recognition accuracy by increasing the number of iterations. However, an increasing number of iterations comes at a high computational cost. This can be prohibitive with large databases or when retraining is required as new data become available. In this study, we used a very large dataset comprising more than 5,000 individually annotated cardiac MRI scans of the same number of subjects, which is 50-fold the 100 cases used in our previous study [271]. However, when compared to natural image datasets [118], our cardiac MRI dataset is still relatively small. We had to design an efficient network taking full advantage of the available data. Considering there were only a few labelled images, there was no point in constructing a network with too many sub-sampling layers; there would have been a higher computational cost with more layers of feature abstraction. Three-dimensional CNN have been among the most promising solutions for object detection tasks. Thus far, most studies have focused on image segmentation and registration, and little effort has been devoted to AIQA. We propose a FD3D CNN with an extra layer using a Fisher discriminant criterion, which tackles the problem of detecting full LV coverage as an important quality criterion. Our method can eliminate redundant convolutional computations during forward propagation and achieve a comparable result with a smaller number of training samples and iterations. Specifically, our FD3D CNN can achieve a high precision rate of nearly 92%/91% for MBS/MAS detection with only 20 epochs, which is better than traditional 3D CNN. Meanwhile, even with a small number of training samples ($4 \times 10{,}000$), our FD3D CNN can decrease the error rate by approximately 29.1% compared to traditional 3D CNN approaches for MBS detection.

Our proposed automatic assessment framework for full LV coverage has great potential to improve the robustness of subsequent population image parsing. One can imagine an approach whereby image analysis is adaptive to image quality and where different models are used depending on whether the volume under analysis is missing basal or apical slices. In our architecture, we focus on learning intensity representations and develop a FD3D CNN

to describe those that best discriminate the missing apical or basal slices. We then use the computed representation classifiers to identify the final image quality. The advantages of a representation-based method for vision tasks are manifold: they can be composed to create descriptions at various levels of specificity; they are generalizable, as they can be learned once and then applied to recognize new objects or categories with no further training and are efficient, possibly requiring exponentially fewer representations than explicitly naming each category. In the future, we plan to investigate the possibility of detecting full LV coverage for all slices, rather than just for basal/apical slices, so we can directly predict visual quality scores. The difficulty of detecting missing middle slices lies in the similar shape of contiguous LV slices, which makes training the representation classifier a non-trivial task. Another future work is to extend deep-learning methods for multi-plane estimation, that is, regressing one 3D volume to estimating missing slices acquired from different positions. This is a limitation of our two-stage framework, which can only estimate the basal and apical planes. One way to achieve 3D CNN for multi-plane estimation would be to apply regression on each plane separately and then combine all regression results into a single estimation.

4.6 Conclusion

In this study, we tackled the problem of detecting incomplete LV coverage in large population image databases. We illustrated the concept by proposing a Fisher discriminative 3D CNN tested on CMR data from the UK Biobank. Our FD3D CNN was proposed by adding a new Fisher-discriminative fully connected layer into the network, which achieved a significant improvement in intensity representation. The learned representation classifiers were computed for candidates of corresponding quality categories. We also validated our model by training with the UK Biobank dataset and cross-evaluating with data from the Data Science Bowl Cardiac Challenge dataset. The proposed model shows high consistency with human perception and is superior to state-of-the-art methods, showing its high potential. Our proposed FD3D CNN can also be easily applied and boosts results for other detection and segmentation tasks in medical image analysis.

Chapter 5

Multi-Input and Dataset-Invariant Adversarial Learning (MDAL) for Left and Right-Ventricular Coverage Estimation in Cardiac MRI

This chapter is based on:

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, Multi-Input and Dataset-Invariant Adversarial Learning (MDAL) for Left and Right-Ventricular Coverage Estimation in Cardiac MRI, *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 481-489, Springer, Cham, 2018.*

Authors' contributions: L.Z., M.P. and A.F.F. conceived and designed the study; S.E.P., S.N. and S.P. provided support on clinical aspects and he also provided the UK Biobank data resource to be used for training and testing; L.Z. designed the method, performed data analysis and wrote the manuscript. All authors read and approved the manuscript.

Abstract: Cardiac functional parameters, such as, the Ejection Fraction (EF) and Cardiac Output (CO) of both ventricles, are most immediate indicators of normal/abnormal cardiac function. To compute these parameters, accurate measurement of ventricular volumes at end-diastole (ED) and end-systole (ES) are required. Accurate volume measurements depend on the correct identification of basal and apical slices in cardiac magnetic resonance (CMR) sequences that provide full coverage of both left (LV) and right (RV) ventricles. This chapter proposes a novel adversarial learning (AL) approach based on convolutional neural networks (CNN) that detects and localizes the basal/apical slices in an image volume independently of image-acquisition parameters, such as, imaging device, magnetic field strength, variations in protocol execution, etc. The proposed model is trained on multiple cohorts of different provenance, and learns image features from different MRI viewing planes to learn the appearance and predict the position of the basal and apical planes. To the best of our knowledge, this is the first work tackling the fully automatic detection and position regression of basal/apical slices in CMR volumes in a dataset-invariant manner. We achieve this by maximizing the ability of a CNN to regress the position of basal/apical slices within a single dataset, while minimizing the ability of a classifier to discriminate image features between different data sources. Our results show superior performance over state-of-the-art methods.

Keywords: Deep Learning · Dataset Invariance · Adversarial Learning · Ventricular Coverage Assessment · MRI.

5.1 Introduction

To obtain accurate and reliable volume and functional parameter measurements in CMR imaging studies, recognizing basal and apical slices for both ventricles is crucial. Unfortunately, current practice to detect basal/or apical slice positions is still carried out by visual inspection of experts on the image. This practice is costly, subjective, error prone, and time consuming [9]. Although significant progress [271] has been made in automatic assessment of full LV coverage in cardiac MRI, to accurately measure volumes and functional parameters for both ventricles where the basal/apical slices are missing, methods to estimate the position of the missing slices are required [178]. Such methods would be critical to prompt the intervention of experts to correct problems in data measurements, or to trigger algorithms that can cope with missing data by, for instance, imputation [75] through image synthesis, or shape based extrapolation. This paves the way to "quality-aware image

analysis" [253]. To the best of our knowledge, previous work regarding image quality control has focused solely on coverage detection of the LV, but not on missing slice position estimation.

In medical image analysis, it is sometimes convenient or necessary to infer an image in one modality from another for image quality assessment purposes. One major challenge of basal/apical slice estimation for CMR comes from differences between data sources, which are tissue appearance and/or spatial resolution of images sourced from different physical acquisition principles or parameters. Such differences make it difficult to generalize algorithms trained on specific datasets to other data sources. This is problematic not only when the source and target datasets are different, but more so, when the target dataset contains no labels. In all such scenarios, it is highly desirable to learn a discriminative classifier or other predictor in the presence of a shift between training and test distributions, which is called *dataset invariance* [70]. The general approach of achieving dataset adaptation has been explored under many facets. Among the existing cross-dataset learning works, dataset adaptation has been adopted for re-identification hoping labeled data from a source dataset can provide transferable identity-discriminative information for a target dataset. [97] explored the possibility of generating multimodal images from single-modality imagery. [134] [151] employed multi-task metric learning models to benefit the target task. However, these works are focused mainly on linear assumptions.

In this chapter, we focus on the non-linear representations and analysis of short-axis (SA) and long-axis (LA) cine MRI for the detection and regression of the basal and apical slices of both ventricles in CMR volumes. To deal with the problem where there is no labeled data for a target dataset, and one hopes to transfer knowledge from a model trained on sufficient labeled data of a source dataset sharing the same feature space, but with a different marginal distribution we present these contributions: 1) We present a unified model (MDAL) for any cross-dataset basal/apical slice estimation problem in CMR volumes; 2) We integrate adversarial feature learning by building an end-to-end architecture of CNNs and transferring non-linear representations from a labeled source dataset to a target dataset where labels are non-existent. Our deep architecture effectively improves the adaptability of learning with data of different databases; 3) A multi-view image extension of the adversarial learning model is proposed and exploited. By making use of multi-view images acquired from short- and long-axis views, one can further improve and constrain the basal/apical slice position. We evaluate our method on three datasets and compare with state-of-the-art methods. Experimental results show the superior performance of our method compared to other approaches.

5.2 Methodology

5.2.1 Problem Formulation

The cross-dataset localization of basal or apical slices can be formulated as two tasks: (*i*) *Dataset Invariance:* given a set of 3D images $\mathscr{X}^s = [\mathbf{X}_1^s, ..., \mathbf{X}_N^s] \in \mathbb{R}^{m \times n \times z^s \times N^s}$ of modality $\mathscr{M}_s$ in the source dataset, and $\mathscr{X}^t = [\mathbf{X}_1^t, ..., \mathbf{X}_N^t] \in \mathbb{R}^{m \times n \times z^t \times N^t}$ of modality $\mathscr{M}_t$ in the target dataset. m, n are the dimensions of axial view of the image, and z^s and z^t denote the size of images along the z-axis, while N^s and N^t are the number of volumes in source and target datasets, respectively. Our goal is to build mappings between the source (training-time) and the target (test-time) datasets, that reduce the difference between the source and target data distributions. An overview of the schematic of our dataset-invariant adversarial network is depicted in Figure 5.1. (*ii*) *Multi-view Slice Regression:* In this task, slice localization performance is enhanced by using multiple image stacks, *e.g.* SA and LA stacks, into a single regression task. Let $\mathbf{X}^s = \{\mathbf{x}_i^s, r_i^s\}_{i=1}^{Z^s}$ and $\mathbf{Y}^s = \{\mathbf{y}_i^s, r_i^s\}_{i=1}^{Z^s}$ be a labeled 3D CMR volume from source modality $\mathscr{M}_s$ in short- and long-axis, respectively, and $\mathbf{x}_b^s$, $\mathbf{x}_a^s$, and $\mathbf{y}_b^s$, $\mathbf{y}_a^s$ be the short-axis slices, and long-axis image patches of the basal and apical views; let $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{Z^t}$ and $\mathbf{Y}^t = \{\mathbf{y}_i^t\}_{i=1}^{Z^t}$ represent an unlabeled sample from the target dataset in short- and long-axis, i represents the i^{th} slice and Z is the total number of CMR slices. Our goal is to learn the discriminative features from $\mathbf{x}_b^s$, $\mathbf{x}_a^s$, and $\mathbf{y}_b^s$, $\mathbf{y}_a^s$ to localize the basal and apical slices in two axes for CMR volumes in the target dataset. We use the labeled UK Biobank (UKBB) [183] cardiac MRI data cohort together with the MESA and DETERMINE datasets, and apply our method to cross-dataset basal and apical slice regression tasks.

5.2.2 Multi-Input and Dataset-Invariant Adversarial Learning

Inspired by Adversarial Learning (AL) [81] and Dataset Adaptation (DA) [213] for cross-dataset transfer, we propose a Dataset-Invariant Adversarial Learning model, which extends the DA formulation into a AL strategy, and performs them jointly in a unified framework. We propose multi-view adversarial learning by creating multiple input channels (MC) from images which are re-sampled to the same spatial grid and visualize the same anatomy. An overview of our method is depicted in Figure 5.2. Given two sets of slices $\{\mathbf{x}_i^s\}_{i=1}^N$, $\{\mathbf{y}_i^s\}_{i=1}^N$ with slice position labels $\{r_i^s\}_{i=1}^N$ for training, to learn a model that

Notation: Matrices and 3D images are written in bold uppercase (*e.g.*, image $\mathbf{X}$, $\mathbf{Y}$), vectors and vectorized 2D images in bold lowercase (*e.g.*, slice $\mathbf{x}$, $\mathbf{y}$) and scalars in lowercase (*e.g.*, slice position label r).

http://www.cardiacatlas.org/studies/mesa/

http://www.cardiacatlas.org/studies/determine/

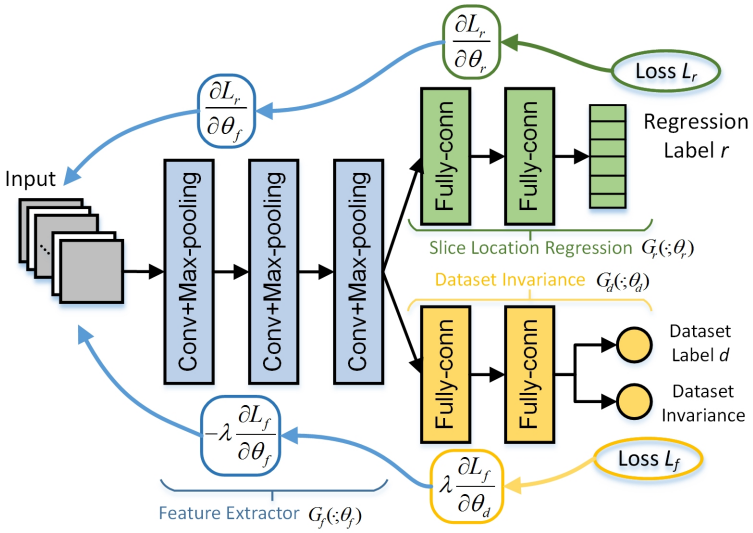


Figure 5.1: Schematic of our dataset-invariant adversarial network.

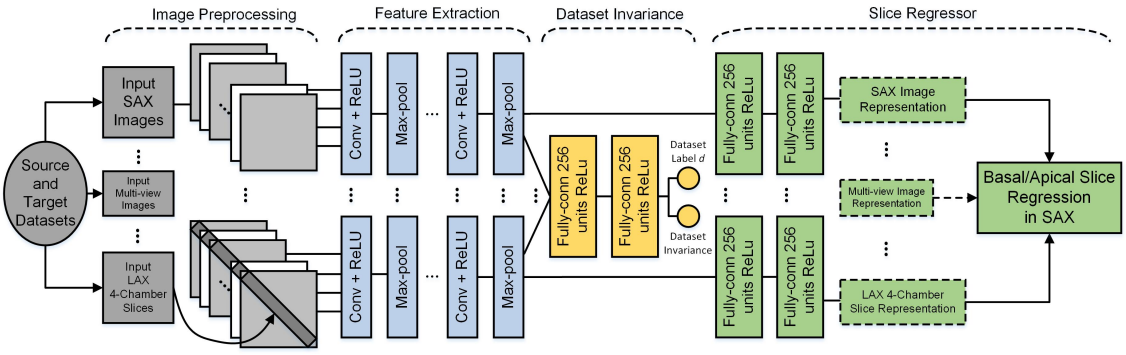


Figure 5.2: System overview of our proposed dataset-invariant adversarial model with multi-view input channels for bi-ventricular coverage estimation in cardiac MRI. Each channel contains three conv layers, three max-pooling layers and two fully-connected layers. Additional dataset invariance net (yellow) includes two fully-connected layers. Kernel numbers in each conv layer are 16, 16 and 64 with sizes of 7×7, 13×13 and 10×10, respectively; filter sizes in each max-pooling layer are 2×2, 3×3 and 2×2 with stride 2.

can generalize well from one dataset to another, and is used both during training and test time to regress the basal/apical slice position, we optimize this objective in stages: 1) we optimize the label regression loss

$$\begin{aligned}\mathscr{L}_r^i &= \mathscr{L}_r(G_{sigm}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s;\theta_f);\theta_r),r_i)\\ &= \sum_i \left\|r_i - G_{sigm}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s;\theta_f);\theta_r),r_i)\right\|_2^2 + \frac{1}{2}\left(\left\|\theta_f\right\|_2^2 + \|\theta_r\|_2^2\right),\end{aligned} \tag{5.1}$$

where θ_f is the representation parameter of the neural network feature extractor, which corresponds to the feature extraction layers. θ_r is the regression parameter of the slice regression net, which corresponds to the regression layers. r_i denotes the i^{th} slice position label. θ_f and θ_r are trained for the i^{th} image by using the labeled source data $\{\mathbf{X}_i^s,\mathbf{r}_i^s\}_{i=1}^{N^s}$ and $\{\mathbf{Y}_i^s,\mathbf{r}_i^s\}_{i=1}^{N^s}$. 2) Since dataset adversarial learning satisfies a dataset adaptation mechanism, we minimize source and target representation distances through alternating *minimax* between two loss functions: one is the dataset discriminator loss

$$\begin{aligned}\mathscr{L}_d^i &= \mathscr{L}_d(G_{disc}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d_i)\\ &= -\sum_i \mathbb{I}[o_d = d_i]\log\big(G_{disc}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d_i\big),\end{aligned} \tag{5.2}$$

which classifies whether an image is drawn from the source or the target dataset. o_d indicates the output of the dataset classifier for the i^{th} image, θ_d is the parameter used for the computation of the dataset prediction output of the network, which corresponds to the dataset invariance layers; d_i denotes the dataset that the example slice i is drawn from. The other is the source and target mapping invariant loss

$$\begin{aligned}\mathscr{L}_f^i &= \mathscr{L}_f(G_{conf}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d_i)\\ &= -\sum_d \frac{1}{D}\log\big(G_{conf}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d_i\big),\end{aligned} \tag{5.3}$$

which is optimized with a constrained adversarial objective by computing the cross entropy between the output predicted dataset labels, and a uniform distribution over dataset labels. D indicates the number of input channels. Our full method then optimizes the joint loss function

$$\begin{aligned}E(\theta_f,\theta_r,\theta_d) = \mathscr{L}_r(G_{sigm}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s;\theta_f);\theta_r),r)\\ +\lambda\mathscr{L}_f(G_{conf}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d),\end{aligned} \tag{5.4}$$

where hyperparameter λ determines how strongly the dataset invariance influences the optimization; $G_{conv}(\cdot)$ is a convolution layer function that maps an example into a new representation; $G_{sigm}(\cdot)$ is a label prediction layer function; $G_{disc}(\cdot)$ and $G_{conf}(\cdot)$ are the dataset prediction and invariance layer functions.

5.2.3 Optimization

Similar to classical CNN learning methods, we propose to tackle the optimization problem with the stochastic gradient procedure, in which updates are made in the opposite direction of the gradient of Equation (5.4) to minimize parameters, and in the direction of the gradient to maximize other parameters [70]. We optimize the objective in the following stages.

Optimizing the Label Regressor: In adversarial adaptive methods, the main goal is to regularize the learning of the source and target mappings, so as to minimize the distance between the empirical source and target mapping distributions. If so then the source regression model can be directly applied to the target representations, eliminating the need to learn a separate target regressor. Training the neural network then leads to this optimization problem on the source dataset:

$$\arg\min_{\theta_f,\theta_r}\{\frac{1}{N^s}\sum_{i=1}^{N^s}\mathscr{L}_r^i(G_{sigm}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s;\theta_f);\theta_r),r_i)\}. \tag{5.5}$$

Optimizing for Dataset Invariance: This optimization corresponds to the true *minimax* objective ($\mathscr{L}_d$ and $\mathscr{L}_f$) for the dataset classifier parameters and the dataset invariant representation. The two losses stand in direct opposition to one another: learning a fully dataset invariant representation means the dataset classifier must do poorly, and learning an effective dataset classifier means that the representation is not dataset invariant. Rather than globally optimizing θ_d and θ_f, we instead perform iterative updates for these two objectives given the fixed parameters from the previous iteration:

$$\arg\min_{\theta_d}\{-\frac{1}{\mathscr{N}}\sum_{i=1}^{\mathscr{N}}\mathscr{L}_d^i(G_{disc}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d_i)\}, \tag{5.6}$$

$$\arg\max_{\theta_f}\{-\frac{1}{\mathscr{N}}\sum_{i=1}^{\mathscr{N}}\mathscr{L}_f^i(G_{conf}(G_{conv}(\mathbf{x}_s,\mathbf{y}_s,\mathbf{x}_t,\mathbf{y}_t;\theta_f);\theta_d),d_i)\}, \tag{5.7}$$

where $\mathscr{N}=N^s+N^t$ being the total number of samples. These losses are readily implemented in standard deep learning frameworks, and after setting learning rates properly so Equation (5.6) only updates θ_d and (5.7) only updates θ_f, the updates can be performed via standard backpropagation. Together, these updates ensure that we learn a representation that is dataset invariant. We summarize the proposed method in the following Algorithm 1.

Algorithm 2: MIDL Algorithm.

Input: samples $S \sim \{(\mathbf{x}_i, \mathbf{y}_i, r_i)\}_{i=1}^{N^s}$ and $T \sim \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N^t}$; adaptation parameter λ; learning rate μ;
Output: Neural network parameters $\{\theta_f, \theta_r, \theta_d\}$;
Initialize: $\theta_f, \theta_y \leftarrow$ random_init; $\theta_d, y_d \leftarrow 0$;
while *stopping criterion has not been met* **do**
  for *i from 1 to* $\mathcal{N}$ **do**
    1) Calculate θ_f and θ_r using (5.5);
    2) Calculate θ_d using (5.6) with fixed θ_f;
    3) Calculate θ_f using (5.7) with fixed θ_d;
    4) Update the parameters using $\theta_f \leftarrow \theta_f - \mu(\frac{\partial L_r^i}{\partial \theta_f} - \lambda \frac{\partial L_d^i}{\partial \theta_f})$, $\theta_r \leftarrow \theta_r - \mu \frac{\partial L_r^i}{\partial \theta_r}$, $\theta_d \leftarrow \theta_d - \mu \frac{\partial L_r^i}{\partial \theta_d}$.
  end
end

5.2.4 Detection and Regression for Basal/Apical Slice Position

We denote $\hat{\mathscr{H}}_t$, $\hat{\mathscr{G}}_t$ as extracted query features, and $\hat{\mathscr{H}}_s$, $\hat{\mathscr{G}}_s$ as extracted basal/apical slice representations from SAX and LAX, respectively. In order to regress basal and apical slices according to query features, we compute the dissimilarity matrix $\delta_{i,j}$ based on $\hat{\mathscr{H}}_t$, $\hat{\mathscr{G}}_t$ and $\hat{\mathscr{H}}_s$, $\hat{\mathscr{G}}_s$ using the volume's inter-slice distance as: $\delta_{i,j}(\hat{\mathscr{H}}_t, \hat{\mathscr{H}}_s, \hat{\mathscr{G}}_t, \hat{\mathscr{G}}_s) = \sqrt{(\hat{\mathscr{H}}_t^i - \hat{\mathscr{H}}_s^j)^2 + (\hat{\mathscr{G}}_t^i - \hat{\mathscr{G}}_s^j)^2}$. Then, ranking can be carried out based on the ascending order of each row of the dissimilarity distance, *i.e.*, the lower the entry value $\delta_{i,j}$ is, the closer the basal/apical slice and the query slice are.

5.3 Experiments and Analysis

Data specifications: Basal slices including the left ventricular outflow tract, pulmonary valve and right atrium, and apical slices with a visible ventricular cavity were labeled manually. The distance between the actual location of the basal/apical slice to other slices in the volume were used as training labels for the regression. We validated the proposed MDAL on three target datasets: UKBB, DETERMINE and MESA (protocols of the three datasets are shown in Table 3.2). To prevent over-fitting due to insufficient target data, and to improve the detection rate of our algorithm, we employ data augmentation techniques to artificially enlarge the target datasets. For this purpose we chose a set of realistic rotations, scaling factors, and corresponding mirror images, and applied them to the MRI images. The set of rotations chosen were $-45°$ and $45°$, and the scaling factors 0.75 and 1.25. This

Table 5.1: Cardiovascular magnetic resonance protocols for UKBB, MESA and DETERMINE Datasets.

Dataset	View	Number of Sequences	Cardiac Phases	Matrix Size	Slice Thickness	Slice Gap	Slice Spacing	Slices per Volume
UKB	SAX	4280	50	208×187	8 mm	2 mm	10 mm	ca. 10
	LAX	4280	50	208×187	6 mm	n.a	n.a	1
MESA	SAX	298	20∼30	256×160	6 mm	4 mm	10 mm	ca. 10
	LAX	298	20∼30	256×160	6 mm	n.a	n.a	1
DETERMINE	SAX	300	25	128×256	$\leqslant$ 10 mm	$\leqslant$ 2 mm	10 mm	ca. 10
	LAX	300	25	128×256	6 mm	n.a	n.a	1

increased the number of training samples by a factor of eight. After data augmentation, we had 2400, and 2384 sequences for DETERMINE and MESA datasets, respectively. For evaluating of multi-view models, we defined two input channels, one for SAX images, and another for LAX (4-chamber) from the UKBB, MESA and DETERMINE. The LAX image information was extracted by collecting pixels values along the intersecting line between the 4-chamber view plane and corresponding short-axis plane over the cardiac cycle. We extracted 4 pixels above and below the two plane intersection. We embedded the constructed profile within a square image with zeros everywhere except the profile diagonal (see Figure 5.2 bottom channel).

Experimental set-up: The architecture of our proposed method is shown in Figure 5.2. To maximize the number of training samples from all datasets, while preventing biased learning of image features from a particular dataset and given that the number of samples from the UKBB is at least an order of magnitude larger than from MESA or DETERMINE, we augmented both the MESA and DETERMINE datasets, to match the resulting number of samples from the UKBB. This way our dataset classification task will not over-fit to anyone sample. Our MDAL method processes images with small blocks (120×120), which are crop-centered on the images to extract specific regions of interest. The experiments here reported were conducted using the ConvNet library [51] on an Intel Xeon E5-1620 v3 @3.50GHz machine running Windows 10 with 32GB RAM and Nvidia Quadro K620 GPU. We optimize the network using a learning rate μ of 0.001 and set the hyper-parameter parameter λ to be 0.01, respectively. To evaluate the detection process, we measure classification accuracy, and to evaluate the regression error between the predicted position and the ground truth, we use the Mean Absolute Error (MAE).

Results: We evaluate the performance of the multi-view basal/apical slice detection

Table 5.2: The comparison of basal/apical slice detection accuracy (Mean ± standard deviation) (%) between adaptation and non-adaptation methods, each with single (SAX)- and multi-view inputs (BS/AS indicate basal/apical slice detection accuracy). Best results are highlighted in bold.

Dataset	No dataset adaptation (BS/AS)		With dataset Adaptation (BS/AS)	
	Single-view [271]	Multi-view [271]	Single-view [70]	Multi-view (Ours)
UKBB	79.0±0.2/76.2±0.3	**89.2**±0.1/**92.4**±0.2	78.2±0.2/75.4±0.3	88.7±0.1/91.4±0.3
MESA	31.6±0.3/35.1±0.1	61.5±0.2/68.3±0.4	74.2±0.2/72.9±0.4	**87.1**±0.3/**90.2**±0.2
DETERMINE	48.3±0.2/51.1±0.3	75.6±0.3/78.4±0.3	77.2±0.3/76.5±0.2	**89.0**±0.2/**91.2**±0.2

Table 5.3: Regression error comparison between adaptation and non-adaptation methods, each with single (SAX)- and multi-view inputs for cardiac SAX slice position regression in terms of MAE (Mean ± standard deviation)(mm)(BS/AS indicate basal/apical slice regression errors). Best results are highlighted in bold.

Dataset	No dataset adaptation (BS/AS)		With dataset adaptation (BS/AS)	
	Single-view [271]	Multi-view [271]	Single-view [70]	Multi-view (Ours)
UKBB	4.32±1.6/5.73±1.9	**3.42**±1.1/**3.98**±1.7	5.13±2.1/6.33±2.3	3.64±1.9/4.02±2.0
MESA	7.78±2.0/8.34±2.4	6.47±1.7/6.83±1.4	4.81±1.0/5.73±1.5	**3.98**±1.1/**4.07**±1.3
DETERMINE	6.43±1.9/6.81±2.0	6.01±1.3/6.17±1.4	4.73±1.6/4.81±1.3	**4.24**±1.0/**4.45**±1.3

and regression tasks with and without dataset invariance (adaptation vs non-adaptation), by transferring object regressors from the UKBB to MESA and DETERMINE. To evaluate performance on MESA and DETERMINE, we manually generated annotations as follows: we checked one slice above and below the detected basal slice to confirm the slice is the basal and record true or false, ditto for apex. We chose the CNN architecture in [271] for single- and multi-view metrics with non-adaptation, and the GTSRB architecture in [70] for single-view adaption method. Table 5.2 shows the detection accuracy for basal/apical slice of the adaptation and non-adaptation from single and multi-view. For both test datasets, the best improvements are the result of combining both of these features. For MESA the detection accuracy was increased by 64%, and for DETERMINE best improvements are of 44% (right-most column). Table 5.3 shows the average regression errors of slice locations in millimeter (*mm*). Even without using the multi-input channels, our dataset invariance framework is able to reduce the slice localization error to less than half the average slice spacing found on our test datasets, *i.e.*, $< 5mm$. With multi-view we reduced the localization errors to 4.24 and 4.45mm on average for both basal/apical slices. All the experiments are significantly different at $p < 0.05$.

5.4 Discussion

Our proposed automatic image quality assessment framework for LV and RV coverage estimation in cardiac cine MRI has great potential for the later robust population image analysis. One could imagine that the image analysis methods are adaptive to image quality and design depending on whether the image under analysis is incomplete ventricle coverage. In our architecture, we focus on learning common representation across datasets and develop a regression network to localize the slice position.

There are diverse advantages of an adversarial learning based representation for visual tasks, some of them are listed as follows: they can create common representations among different specific datasets; they are generalizable since they only need to learn all datasets once and then, apply them to identify new objects or categories without further training. One of our future studies is to investigate the possibility of quantifying the ventricle coverage with slice position and correct plane orientation, not specific for ventricle coverage estimation. It is difficult to calculate the percentage of ventricle coverage accurately with incorrect slice orientations and thus, training the volume classifier could be a non-trivial task because of the different slice orientations. Another future study is to extend the deep learning method for synthesizing the slices with correct orientation. One possible way to achieve deep learning approach for image synthesis would be to apply generative adversarial network and synthesize the standard cardiac plane using the adversarial approach on UKBB.

5.5 Conclusion

In this chapter, we have proposed a Multi-Input and Dataset-Invariant Adversarial Learning (MDAL) framework capable of learning a common image representation, and using it to detect and localize basal and apical CMR slices, we achieve this by: first, using a Dataset-Invariant Adversarial Learning (DIAL) model to fit the joint distribution over the images from different datasets with a minimax game. Second, extending the DIAL model to handle multiple view input scenarios thereby obtaining better results for Left and Right-Ventricular coverage estimation in Cardiac MRI. And third, by introducing a regressor network able to predict the location of basal/apical slices. We evaluated our framework on two large datasets MESA and DETERMINE and found that our approach significantly outperforms state-of-the-art non-dataset-adaptive and single-input methods. Finally, Our MDAL framework can be easily generalized to any anatomical structure or image modality.

Chapter 6

Automatic Plane Pose Estimation Across Cardiac Cine MRI Datasets via Deep Adversarial Ranking Nets with Privileged Information

This chapter is based on:

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Automatic Plane Pose Estimation Across Cardiac Cine MRI Datasets via Deep Adversarial Ranking Nets with Privileged Information*, Submitted.

Authors' contributions: L.Z., M.P. and A.F.F. conceived and designed the study; S.E.P., S.N. and S.P. provided support on clinical aspects and he also provided the UK Biobank data resource to be used for training and testing; L.Z. designed the method, performed data analysis and wrote the manuscript. All authors read and approved the manuscript.

6.1 Introduction

Cardiac Magnetic Resonance (CMR) imaging is the reference standard imaging technique used to evaluate morphology and functionality of the heart. After acquisition, automatic techniques can extract volumetric information and derive clinical indexes that place the subject within predetermined population ranges of normality. CMR image acquisition is for the most part automatic, except for the initial localization and framing of the heart done by a trained radiologist or image technician. Because the heart is a moving organ, and the length of the procedure requires the patient to hold their breath multiple times during the exam, resulting images may suffer from artifacts due to variability in the breath-hold position adopted by the patient during each breath-hold. If the initial framing of the heart does not allow a sufficient margin around the organ, these differences in breath-hold may cause the heart to move out of frame, resulting in incomplete coverage either at the basal, or apical region of the organ. A related source of organ coverage variability is the determination of what constitutes a sufficient margin around the heart. Though anatomical features allow the precise localization of the base and apex of the heart, a "sufficient" margin above and below base and apex may not be as precisely defined. This means that slightly different practices may be in place at different imaging facilities or by different experts, resulting in image volumes that while providing full coverage, some may present with one image slice above/below the cardiac base/apex, while others with two image slices above/below the cardiac base/apex. These variations in procedure may present problems for subsequent image analysis algorithms trained under the assumption of consistent object coverage.

Related to object coverage, is consistent orientation of image planes regarding the cardiac ventricles, where, if the slice orientation deviates significantly from expected values, local image structure may change enough to cause subsequent image feature-based algorithms to fail in localizing key features required for further morphological and functional analysis.

These sources of variability may affect the subsequent application of automatic methods for the computation of tissue volumes, cardiovascular indexes, and statistics derived from them. Typically, volume computations are performed on the output of image segmentation algorithms. Whether these algorithms are generative or discriminative, having incomplete or overcomplete organ coverage may cause incorrect segmentations leading to biased estimation of volume parameters. For example, 2D discriminative CNN-based segmentation methods will grossly under/over-estimate blood volumes with incomplete/overcomplete image stacks. 3D generative-based models, such as ASMs, can handle incomplete/overcomplete volumes as they are constrained by shape priors, however, the

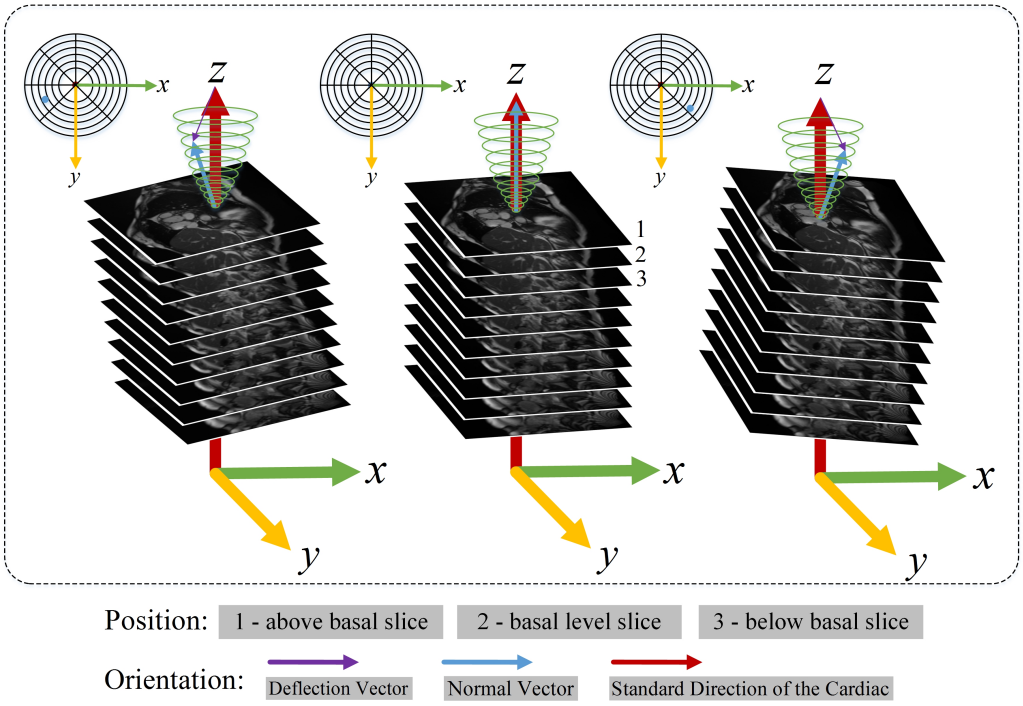


Figure 6.1: Potential issues affecting CMR image acquisitions. In the right and left volumes, short axis slices are acquired with incorrect orientations; In the middle volume, different slices show different position compared with basal slice. Best viewed in color.

performance of ASMs can be affected by incorrect model initialization if the initial shape estimate is poorly positioned regarding the target structure, the algorithm may not converge to the best solution.

Algorithms to robustly determine the expected location and pose parameters of the basal and apical slices in CMR are necessary. Current practice to optimize CMR slice pose is carried out manually, which can be tedious, subjective, error prone, and time consuming [9]. Although significant progress [271] has been made in automatic assessment of full LV coverage in cardiac cine MRI, to accurately measure volumes for both ventricles from volumes with missing basal/or apical slices, the missing slice position and orientation in short-axis needs to be specified [178] to intervene and correct problems in the data soon. Such corrections may include imputation of missing data, choosing the appropriate segmentation method, or excluding faulty image volumes from being used in the computation of aggregated statistics over large patient cohorts [75]. This paves the way to "quality-aware image analysis" [253].

To the best of our knowledge, this is the first study tackling the problem of estimating the slice pose for both ventricles in cardiac MRI. The main contributions of this work are summarized:

1. This is the first study to exploit deep learning for automatic CMR slice pose estimation. The CNN sufficiently encodes spatial contextual image information and hierarchically extracts high-level features in a data driven way.

2. To efficiently leverage CNN, we present an adversarial model (DIAL) to process any cross-data set problems in cardiac MRI, which allows training based on annotated data in the source data set, and test on un-annotated data in the target data set.

3. We propose an end-to-end MLMT regression network to jointly optimize slice pose estimation with a set of related tasks (distance and orientation). Our proposed MLMT model has great generalization capability and works well on data with different labels.

4. We propose and formulate a new problem, which combines the DIAL and MLMT models with a novel PI loss. We call the proposed method data set Adversarial Regression Network with PI (DARN*). To the best of our knowledge, this is the first work exploiting PI in cardiac MRI instead of using multi-view inputs and it is a much more practical approach in real world applications.

5. We quantitatively assess the performance of our technique in three large-scale 3D cardiac MRI databases achieving comparable results.

In Section 6.2, we review related work. In Section 6.3, we define the LVRV slice pose estimation problem and introduce our proposed method. Experiments, results and discussion of the proposed method are drawn in Sections 6.4 and 6.5, and we conclude the study in Section 6.6.

6.2 Related Work

Recent work has been published automating LV apex/base detection [274], but no existing research has proposed slice pose estimation based on both cardiac ventricles. Paknezhad et al. [178], for example, proposed an automatic tool that uses the horizontal long-axis (HLA) view to find the basal slice. The basal slice was detected using temporal binary profiles created for each short-axis slice from the segmented HLA slice. Drawbacks of this technique are its dependency on correct segmentation, and existence of the HLA slice. Mahapatra et al. [154] proposes a learning-based method that trains a random forest classifier by extracting intensity, texture, and contextual features from a bounding box

around the annotated points at both sides of the mitral valve. Lu et al. [146] propose another learning-based method by introducing auxiliary markers along with the landmarks to collect more contextual information from the image and help landmark detection. Utilizing such methods for basal slice selection comes from the assumption that the basal slice is the first short-axis slice below the line connecting the mitral valve points.

Real-world images can often be annotated with multiple labels, because an image normally abounds with rich semantic information, such as objects, parts, scenes, actions, and their interactions or attributes. CNNs have been trained for single-label image regression tasks such as age estimation [33], registration [159] and depth prediction [137]. In the field of medical image analysis, Kong et al. [115] proposed a temporal regression network (TempReg-Net) to accurately identify End-Diastole (ED) and End-Systole (ES) frames from MRI sequences. Spampinato et al. [222] proposed a CNN based BoNet architecture for bone age regression. These deep regression models are not suitable for our problem because they cannot process multiple regression tasks for a single image with different labels at the same time in an end-to-end learning framework. Deep multi-label learning (MLL) has been widely used in various computer vision problems, recent studies have proposed MLL approaches [250] to solve image classification problems. In this study, we propose a multi-label multi-task (MLMT) approach for cardiac MRI slice distance and orientation estimation. Unlike the metric regression approaches that treat the single label in an individual task as a proportional quantity, the MLMT regression approach can learn a shared representation to predict all the factors with multi-labels from one image.

The general approach of achieving database adaptation has been explored under many facets. Over the years, a large part of the literature has focused mainly on linear hypothesis. Recently, some research [133], [258] has shown that non-linear neural networks can also be successful at learning features in a data-driven way achieving promising and stable results across domain changes, and can thus be applied to cross-domain transfer. Among them, adversarial learning has been explored for cross-domain tasks, which choose an adversarial loss to minimize domain shift, learning a representation that is simultaneously discriminative of source labels while not being able to distinguish between domains. [240] proposed adding a domain classifier that predicts the binary domain label of the inputs and designed a domain confusion loss to encourage its prediction to be as close as possible to a uniform distribution over binary labels. The gradient reversal algorithm (ReverseGrad) proposed in [70] also treats domain invariance as a binary classification problem, but directly maximizes the loss of the domain classifier by reversing its gradients. Motivated by these works, we propose an adversarial learning approach for dataset adaptation, which seeks to directly map source labeled images onto target unlabeled images.

Data-driven approaches leverage large amounts of training data to determine the optimal model parameters in a bottom-up fashion. Purely data-driven methods are often brittle and prone to fail when learning with limited training data, due to over-fitting or an optimization obstacle involved. In many applications, additional information is often available in the training phase. Vapnik and Vashist [246] referred to such additional information as privileged information (PI) and showed that PI can be utilized as a "teacher" to train more effective models in traditional supervised learning problems. Recently, [155] presented a regularized RNNs with additional information for RGB video sequences. This motivates us to incorporate PI into our DARN model for LVRV slice pose estimation.

6.3 Methodology

6.3.1 Problem Formulation

We formulate our problem as two tasks:

1) *Datasets invariance:* given a set of 3D images $\mathscr{X}^s = [\mathbf{X}_1^s, ..., \mathbf{X}_N^s] \in \mathbb{R}^{m \times n \times z^s \times N^s}$ and corresponding labels $\mathscr{Y}^s = [\mathbf{Y}_1^s, ..., \mathbf{Y}_N^s]$ of modality $\mathscr{M}_s$ in the source dataset, and $\mathscr{X}^t = [\mathbf{X}_1^t, ..., \mathbf{X}_N^t] \in \mathbb{R}^{m \times n \times z^t \times N^t}$ of modality $\mathscr{M}_t$ in the target dataset. m, n are the dimensions of axial view of the image, and z^s and z^t denotes the size of image along the z-axis, while N^s and N^t are the number of elements in source and target datasets, respectively. Let $\{\mathbf{X}^s, \mathbf{Y}\} = \{\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^{Z^s}$ and $\{\mathbf{X}^{*s}, \mathbf{Y}\} = \{\mathbf{x}_i^{*s}, \mathbf{y}_i^s\}_{i=1}^{Z^s}$ be a labeled 3D CMR volume from source modality $\mathscr{M}_s$ in short- and long-axis, respectively; let $\mathbf{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{Z^t}$ represent an unlabeled sample from the target dataset in short-axis, Z is the total number of CMR slices. Our goal is to build mappings between the source (training-time) and the target (test-time) datasets, so that reducing the difference between the source and target dataset distributions;

2) *Slice pose estimation:* In this task, the performance of slice pose estimation is enhanced by using multi-label multitask (MLMT) learning, *e.g.* distance regression task and orientation regression task, into a single deep regression neural network. In the context of MLMT learning, assume there is an image sequence denoted by $\{\mathbf{X}, \mathbf{Y}\} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{Z}$, where each $\mathbf{X}$ has Z slices and $\mathbf{y}$ contains the labels (distance and orientation) associated with $\mathbf{x}$. We represent $\mathbf{y}_i$ as a vector of length C, where C is the number of labels. For example, the k^{th} dimension $\mathbf{y}_i(k)$ denotes the distance between basal slice and the i^{th} slice. Our goal is the training of a regression network processed with MLMT procedure, mapping

Notation: Matrices and 3D images are written in bold uppercase (*e.g.*, image $\mathbf{X}, \mathbf{Y}$), vectors and vectorized 2D images in bold lowercase (*e.g.*, slice $\mathbf{x}, \mathbf{y}$), and scalars are noted in lowercase (*e.g.*, slice position label r).

from images to corresponding probabilities by the function $\eta(\cdot) : \mathscr{X} \rightarrow \mathscr{Y}$. We incorporate the long-axis patches $\mathbf{X}^{*s}$ as privileged information (PI) into the learning system at training time and the testing stage continues to make use of only to $\mathbf{X}^{t}$ without any access to $\mathbf{X}^{*t}$. This training process produces a ConvNet with learnt parameters $\mathbf{W}$ that is effectively a mapping between the input images $\mathscr{X}$, $\mathscr{X}^{*}$ and the estimated output vector $\mathscr{Y}$, represented by

$$\mathscr{Y} = \eta(\mathscr{X}, \mathscr{X}^{*}; \mathbf{W}). \tag{6.1}$$

We use the annotated UK Biobank (UKBB) [183] cardiac MRI data cohort together with the MESA[1] and DETERMINE[2] datasets, and apply our method to cross-dataset slice position and orientation regression tasks. We describe the deep adversarial learning approaches and configure MLMT regression network incorporated with PI to perform automated slice pose estimation across datasets. At the training time, DIAL promotes the emergence of features that are indiscriminate with respect to the shift between the datasets, but discriminative for the main learning task on the source dataset. Instead of using metric regression to identify the slice poses, the MLMT is trained to regress the distance and orientation for each slice at the same time with the aim of mutual benefit. Our goal is to learn the discriminative features from $\mathbf{x}_i^s$, and utilize $\mathbf{x}_i^{*s}$ as PI to train more effective models and estimate the slice poses in short axis for CMR volumes in the target dataset.

6.3.2 Deep Adversarial Learning for Dataset-Invariant

Inspired by Adversarial Learning (AL) and Dataset Adaptation (DA) for cross-dataset transfer, we propose a Dataset-Invariant Adversarial Learning (DIAL) model, which extends the DA formulation into a AL strategy, and performs them jointly in a unified framework. We propose multi-view adversarial learning by creating multiple input channels (MC) from images , which are re-sampled to the same spatial grid and visualize the same anatomy. An overview of our method is depicted in Fig. 6.1. Given a set of slices $\{\mathbf{x}_i^s\}_{i=1}^{Z}$ with corresponding labels $\{\mathbf{y}_i^s\}_{i=1}^{Z}$ for training, to learn a model that can generalize well from one dataset to another, and is used both during training and test time to regress the basal/apical slice pose, we optimize this objective in stages: 1) we optimize the label regression loss

$$\arg\min_{\mathbf{w}_f, \mathbf{w}_y^t} \{\frac{1}{N^s} \sum_{i=1}^{N^s} \mathscr{L}_y^i(G_{sigm}(G_{conv}(\mathbf{x}_i^s; \mathbf{w}_f); \mathbf{w}_y^t), y_i^t)\} \tag{6.2}$$

where $\mathbf{w}_f$ is the representation parameter of the neural network feature extractor, which corresponds to the feature extraction layers. $\mathbf{w}_y^t$ is the regression parameter of the slice

[1] http://www.cardiacatlas.org/studies/mesa/

[2] http://www.cardiacatlas.org/studies/determine/

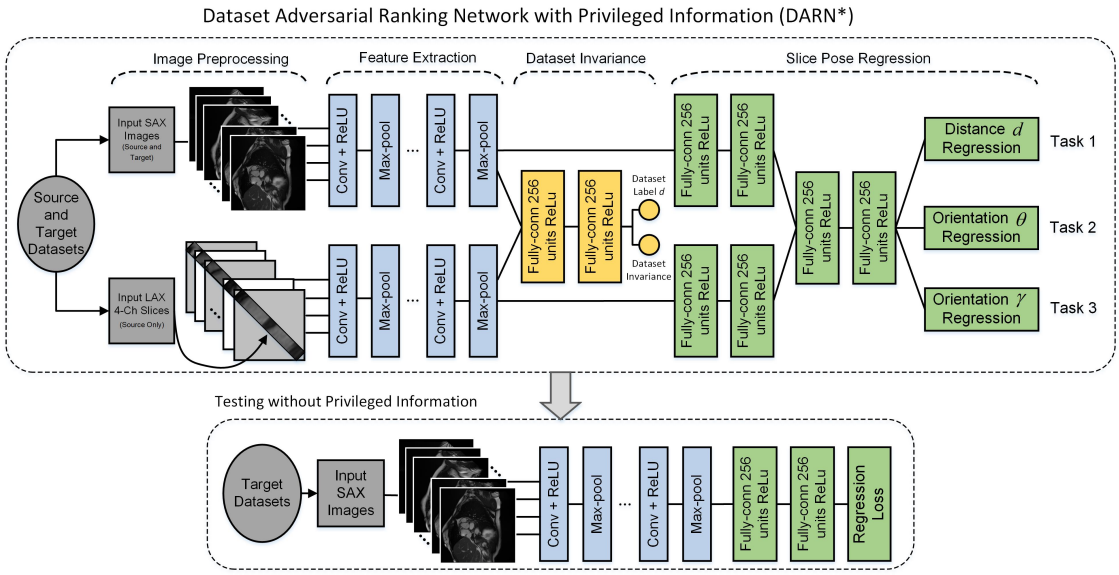


Figure 6.2: The proposed framework of PI-based DARN. Our approach consists of three steps: 1) The CNN acts as a feature extractor to extract the spatial pattern of the cardiac image volume to facilitate the dataset invariance phase; 2) We use a Dataset-Invariant Adversarial Learning (DIAL) model to fit the joint distribution over the images from different datasets with a minimax game; 3) We extend the DIAL model to handle MTRN model with LUPI scenarios. The joint network can be trained to learn the complex spatial patterns of the cardiac sequences cross different CMRI datasets, and give predictions for the slice pose without any privileged information (PI) during testing. Best viewed in color.

regression net, which corresponds to the regression layers. y_i^t denotes the i^{th} slice position label. $\mathbf{w}_f$ and $\mathbf{w}_y^t$ are trained for the i^{th} image by using the labeled source data $\{\mathbf{x}_i^s, \mathbf{y}_i^s\}_{i=1}^{Z^s}$.

2) Since dataset adversarial learning satisfies a dataset adaptation mechanism, we minimize source and target representation distances through alternating *minimax* between two loss functions: one is the dataset discriminator loss

$$\arg\min_{\mathbf{w}_d}\{-\frac{1}{\mathscr{N}}\sum_{i=1}^{\mathscr{N}}\mathscr{L}_d^i(G_{soft}(G_{conv}(\mathbf{x}_i^s,\mathbf{x}_i^t,;\mathbf{w}_f);\mathbf{w}_d),d_i)\} \tag{6.3}$$

which classifies whether an image is drawn from the source or the target dataset. o_d indicates the output of the dataset classifier for the i^{th} image, $\mathbf{w}_d$ is the parameter used for the computation of the dataset prediction output of the network, which corresponds to the dataset invariance layers; d_i denotes the dataset that the example slice i is drawn from. The other is the source and target mapping invariant loss

$$\arg\max_{\mathbf{w}_f}\{-\frac{1}{\mathscr{N}}\sum_{i=1}^{\mathscr{N}}\mathscr{L}_f^i(G_{soft}(G_{conv}(\mathbf{x}_i^s,\mathbf{x}_i^t;\mathbf{w}_d);\mathbf{w}_f),d_i)\} \tag{6.4}$$

which is optimized with a constrained adversarial objective by computing the cross entropy between the output predicted dataset labels, and a uniform distribution over dataset labels. $\mathcal{N} = N^s + N^t$ being the total number of samples, D indicates the number of input channels. Our full method then optimizes the joint loss function

$$\begin{aligned} E(\mathbf{w}_f, \mathbf{w}_d, \mathbf{w}_y^t) = \mathscr{L}_y(G_{sigm}(G_{conv}(\mathbf{x}^s; \mathbf{w}_f); \mathbf{w}_y^t), y^t) \\ + \lambda \mathscr{L}_f(G_{soft}(G_{conv}(\mathbf{x}^s, \mathbf{x}^t; \mathbf{w}_f); \mathbf{w}_d), d), \end{aligned} \tag{6.5}$$

where hyperparameter λ determines how strongly the dataset invariance influences the optimization; $G_{conv}(\cdot)$ is a convolution layer function (feature extraction) that maps an example into a new representation; $G_{sigm}(\cdot)$ is a label prediction (sigmoid) layer function; $G_{soft}(\cdot)$ is a dataset prediction (softmax) layer function.

Similar to classical CNN learning methods, we propose to tackle the optimization problem with the stochastic gradient procedure, in which updates are made in the opposite direction of the gradient of Equation. (6.5) to minimize parameters, and in the direction of the gradient to maximize other parameters [70].

6.3.3 MLMT Learning with Privileged Information

In this section, we provide the learning procedure of our MLMT* learning network for slice position and orientation estimation, instead of using metric regression network. Then, we show that our framework can be trained end-to-end by optimizing the regression and spatial structured constraints.

1) Multi-label Multi-task Learning (MLMT): To fully capture the spatial information relevant to the left- and right-ventricle in every slice, we employ a CNN as the feature extractor in order to efficiently encode the spatial information. We choose DIAL model to extract the features of each slice cross cardiac datasets and predict the corresponding position and orientation with multiple regression tasks.

The traditional multi-task learning (MTL) [269] [277] seeks to improve the generalization performance of multiple related tasks by learning them jointly. Suppose we have a total of T tasks and the training data for the t^{th} task are denoted as, $(\mathbf{x}_i^t, y_i^t)$, where $t = \{1, ..., T\}$, with $\mathbf{x}_i, y_i \in \{\{\mathbf{x}_i^n, y_i^n\}_{i=1}^Z\}_{n=1}^{N^s}$ being the input image and label, respectively. The goal of the MTL is to minimize

$$\underset{\{\mathbf{w}^t\}_{t=1}^T}{\arg\min} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathscr{L}(G_{sigm}(G_{conv}(\mathbf{x}_i; \mathbf{w}_f^t); \mathbf{w}_y^t), y_i^t) + \Phi(\mathbf{w}^t) \tag{6.6}$$

where $\mathbf{w}^t = \{\mathbf{w}_f^t, \mathbf{w}_y^t\}$ is the weight vector for t^{th} task and y_i^t is the label for i^{th} image for t^{th} task. The loss function is denoted by $\mathscr{L}(\cdot)$. A typical choice is the mean square for

regression and the cross-entropy loss for classification. $\Phi(\mathbf{w}^t)$ is the regularization term that penalizes the complexity of weights.

In this work, we divided all the tasks into two groups: regression tasks t_d for distance parameter d and regression tasks t_o for orientation parameters: θ and γ (the definition will be discussed in experiment section). We follow [172] and adopt *data specific scheme* for each task in the two groups, which obtains the distribution of sample number over their distance and orientation, and set the importance parameters according to this distribution. For our MLMT with multiple output, each of them corresponding to regression task for i^{th} image. Let α_y denotes the importance coefficient of the label y ($y \in \{y_d, y_o\}$) in regression tasks. In our approach, the importance parameters are set according to the reliability of different regression parameters. In other words, we set $\alpha_y^{t_d} = \sqrt{N_d}/(\sum_{d=1}^{D}\sqrt{N_d})$ for the distance regression task and $\alpha_y^{t_o} = \sqrt{N_o}/(\sum_{o=1}^{O}\sqrt{N_o})$ for the orientation regression task, where N_d is the number of samples with distance label d, N_o is the number of samples with orientation label o. In particular, the t_d corresponds to a distance regression task, which is trained to regress the slice distance d in a sequence. Thus, for the task t_d, the number of samples with distances nearby d, *e.g.*, samples with distance $\{(d-\delta d), d, (d+\delta d)\}$, ($\delta d \in \{0, 1/Z\}$) is more important than other samples for the training of the task t_d output. In other words, if more samples are with distance close to d, we could better train the corresponding position features, and hence it is better to give a relatively larger importance to it. To this end, the loss function of our regressioning network can be formulated as

$$\begin{aligned}
&\arg\min_{\{\mathbf{w}^t\}_{t=1}^{T}} \sum_{t=1}^{T}\sum_{i=1}^{N} \alpha_y \beta_t \mathscr{L}(G_{sigm}(G_{conv}(\mathbf{x}_i;\mathbf{w}_f^t);\mathbf{w}_y^t), y_i^t) = \\
&\arg\min_{\{\mathbf{w}^{t_d}\}_{t_d=1}^{T_d}, t_d=1} \sum_{t_d=1}^{T_d}\sum_{i=1}^{N} \alpha_y^{t_d} \beta_{t_d} \mathscr{L}(G_{sigm}(G_{conv}(\mathbf{x}_i;\mathbf{w}_f^{t_d});\mathbf{w}_y^{t_d}), y_i^{t_d}) + \\
&\arg\min_{\{\mathbf{w}^{t_o}\}_{t_o=1}^{T_o}} \sum_{t_o=1}^{T_o}\sum_{i=1}^{N} \alpha_y^{t_o} \beta_{t_o} \mathscr{L}(G_{sigm}(G_{conv}(\mathbf{x}_i;\mathbf{w}_f^{t_o});\mathbf{w}_y^{t_o}), y_i^{t_o})
\end{aligned} \tag{6.7}$$

where T_d and T_o indicate the total number of regression tasks for distance and orientation, respective and $T = T_d + T_o$; β_t denotes the importance coefficient of t^{th} task's error; the regularization terms are omitted for simplification.

2) MLMT with Privileged Information (MLMT):* In many image processing tasks, there often exists additional information can help us learn a better model in the training stage. We call this kind of information as privileged information (PI), such as image captions. In other words, the PI provide much more the correct information during training, but in the test stage the model operates without the supervision of the PI. This paradigm is

called Learning Using Privileged Information (LUPI) and was introduced by Vapnik and Vashist [246]. In our model, we construct a two-stream framework, which train the first stream model for SAX images, and the second stream model is trained for the PI (LAX patches). With this configuration, our framework can not only effectively utilizes privileged LAX patches, but also can deal with different types of data flexibly, such as the metadata.

Furthermore, we need a PI loss to replace the original MLMT loss in the training phase, so that we can use the PI as a "teacher" to train a more effective model. We propose to utilize PI to model the loss of training data, penalize the difference of PI modeled loss and true loss, and add the difference as a regularization term to Equation. (6.7). Specifically, assume that for each training SAX image $\mathbf{x}_i$, we have a privileged LAX patch $\mathbf{x}_i^*$. We use a second stream of network (called MLMT-PI) to model PI. Compared to the first stream of network, which models the training SAX images, the goal of the second stream is not to learn a regression model, but to model the loss of the first stream. Denote the output of the second stream for an input privileged patch $\mathbf{x}_i^*$ as $f^*(\mathbf{x}_i^*)$, the two streams share the same loss layer defined by

$$\begin{aligned} &\underset{\{\mathbf{w}^t,\mathbf{w}_f^{*,t}\}_{t=1}^T}{\arg\min} \sum_{t=1}^{T}\sum_{i=1}^{N} \alpha_y\beta_t\mathscr{L}(G_{sigm}(G_{conv}(\mathbf{x}_i;\mathbf{w}_f^t);\mathbf{w}_y^t),y_i^t)+ \\ &\gamma\left\|\alpha_y\beta_t\mathscr{L}(G_{sigm}(G_{conv}(\mathbf{x}_i;\mathbf{w}_f^t);\mathbf{w}_y^t),y_i^t) - G_{conv}^*(\mathbf{x}_i^*;\mathbf{w}_f^{*,t})\right\|_2^2 \end{aligned} \tag{6.8}$$

where $\mathbf{x}_i^{*,t}$ is the i^{th} privileged patch and parameterized by the weight vector $\mathbf{w}_f^{*,t}$, $\|\cdot\|_2^2$ is the L2 norm. Our main hyperprameter is the tradeoff parameter γ, which is tune by cross-validation in a small subset of the training data.

The proposed MLMT with PI can be optimized in an alternating fashion. Specifically, we update the main stream while fixing the parameters of the privileged stream until it converges, and subsequently update privileged stream while fixing the parameters of main stream. This process is repeated for several times until the whole system converges.

In the following, we formulate our fully fledged DARN* model based on Equation. (6.5) and (6.9). Suppose we have a set of feature vectors in a shared feature space across tasks $\{\mathbf{x}_i\}_{i=1}^Z$ and their corresponding labels $\{y_i^d, y_i^{t_d}, y_i^{t_\theta}, y_i^{t_\gamma}\}_{i=1}^Z$, where y_i^d is the target of dataset invariance and the remaining are the targets of slice pose regression, including inferences of 'distance' and 'orientation'. More specifically, $y_i^d \in \{0,1\}$ is binary dataset, $y_i^{t_d}$, $y_i^{t_\theta}$ and $y_i^{t_\gamma}$ are multiple values that represent the distances and orientations in 3D space. It is reasonable to employ the least square and cross-entropy as the loss functions for the main

Algorithm 3: DIAL Algorithm.

Input: source data $\{\mathbf{x}_i^s\}_{i=1}^{Z^s}$, $\{\mathbf{x}_i^{*s}\}_{i=1}^{Z^s}$;
target data $\{\mathbf{x}_i^t\}_{i=1}^{Z^t}$, $\{\mathbf{x}_i^{*t}\}_{i=1}^{Z^t}$;
ground-truth dataset label y_i^d.
Initialize: $\mathbf{w}_f, \mathbf{w}_d, \mathbf{w}_y^d \leftarrow$ random_init; $d \leftarrow 0$;
while *stopping criterion has not been met* **do**
  for *i from 1 to* N^s **do**
    1) Calculate $\mathbf{w}_f$, $\mathbf{w}_y^d$ using Eq.(6.2);
    2) Calculate $\mathbf{w}_d$ using Eq.(6.3) with fixed $\mathbf{w}_f$;
    3) Calculate $\mathbf{w}_f$ using Eq.(6.4) with fixed $\mathbf{w}_d$;
    4) Update the parameters using gradient descent method [70].
  end
end
Output: Neural network parameters $\{\mathbf{w}_f, \mathbf{w}_d, \mathbf{w}_y^d\}$;

task (regression) and the dataset invariance task (classification), respectively. Therefore, the objective function can be rewritten as

$$\begin{aligned}
&\underset{\mathbf{w}^d,\{\mathbf{w}^t,\mathbf{w}^{*,t}\}_{t=1}^T}{\arg\min} \sum_{t=1}^{T}\sum_{i=1}^{N}\frac{1}{2}\alpha_y\beta_t\left\|y_i^t - f(\mathbf{x}_i;\mathbf{w}^t)\right\|_2^2 \\
&+\gamma\left\|\frac{1}{2}\alpha_y\beta_t\left\|y_i^t - f(\mathbf{x}_i;\mathbf{w}^t)\right\|^2 - f(\mathbf{x}_i^*;\mathbf{w}^{*,t})\right\|_2^2 \\
&-\sum_{i=1}^{N}\lambda y_i^d \log\Big(p(y_i^d\,\Big|\,\mathbf{x}_i;\mathbf{w}^d)\Big) + \sum_{t=1}^{T}\left(\left\|\mathbf{w}^d\right\|_2^2 + \left\|\mathbf{w}^t\right\|_2^2\right),
\end{aligned} \tag{6.9}$$

where $f(\mathbf{x}_i;\mathbf{w}^t) = (\mathbf{w}^t)^\top\mathbf{x}_i$ is a linear function. $p(y_i^d = m\,|\mathbf{x}_i) = \exp\{(\mathbf{w}_m^d)^\top\mathbf{x}_i\}\Big/\sum_j\exp\{(\mathbf{w}_j^d)^\top\mathbf{x}_i\}$ is a softmax function, which models the class posterior probability ($\mathbf{w}_j^d$ denotes the j^{th} column of the matrix). In this work, we adopt the CNN to jointly learn the share feature space $\mathbf{x}$, since the unique structure of CNN allows for multitask and shared representation.

6.3.4 Model Implementation

1) Network Structure: As shown in Fig. 6.2, our network consists of three parts: Feature Extraction, Dataset Invariance Learning (DIAL) and Pose Regression (MLMT*). Feature extraction includes three 5×5 convolutional layers (C1, C2 and C3), each followed by a 2×2 max-pooling layers (P1, P2 and P3) with stride 2. Followed by P3, there are two branches: one is DIAL, which consists of two fully-connected layers (F1 and F2), each with 256 Rectified Linear Unit (ReLU) activations neurons; the other one is MLMT*, which includes two fully-connected layers (F3 and F4). The fully connected layers (F5, F6) following two

Algorithm 4: MLMT* Algorithm.

Input: training data $\{\mathbf{x}_i^s\}_{i=1}^{Z^s}$, $\{\mathbf{x}_i^{*s}\}_{i=1}^{Z^s}$;
testing data $\{\mathbf{x}_i^t\}_{i=1}^{Z^t}$;
ground-truth distance label y_i^{td};
ground-truth orientation label $y_i^{t\theta}$ and $y_i^{t\gamma}$.

Initialize: $\mathbf{w}_f, \mathbf{w}_y^t \leftarrow$ random_init;

while *stopping criterion has not been met* **do**

  for *i from 1 to* $\mathcal{N}$ **do**

    1) Calculate $\mathbf{w}_f$ and $\mathbf{w}_y^t$ using Eq.(6.7);

    2) Calculate $\mathbf{w}^t$ and $\mathbf{w}_f^{*,t}$ with PI using Eq.(6.8);

    3) Update the parameters using gradient descent method [70].

  end

end

Output: Neural network parameters and predicted regressor for testing images $\{\eta(\mathbf{x}_i; \mathbf{w}_f, \mathbf{w}_y^t)\}_{i=1}^{Z}$;

Table 6.1: Cardiovascular magnetic resonance protocols for UKBB, MESA and DETERMINE Datasets.

Dataset	View	Number of Sequences	Cardiac Phases	Matrix Size	Slice Thickness	Slice Gap	Slice Spacing	Slices per Volume
UKB	SAX	4280	50	208×187	8 mm	2 mm	10 mm	ca. 10
	LAX	4280	50	208×187	6 mm	n.a	n.a	1
MESA	SAX	298	20∼30	256×160	6 mm	4 mm	10 mm	ca. 10
	LAX	298	20∼30	256×160	6 mm	n.a	n.a	1
DETERMINE	SAX	300	25	128×256	$\leqslant$ 10 mm	$\leqslant$ 2 mm	10 mm	ca. 10
	LAX	300	25	128×256	6 mm	n.a	n.a	1

streams produces a feature vector , which is shared by the multiple tasks in the estimation stage.

2) Implementation Considerations: The experiments here reported were conducted using the ConvNet library [51] on an Intel Xeon E5-1620 v3 @3.50GHz machine running Windows 10 with 32GB RAM and Nvidia Quadro K620 GPU. Networks were optimized using gradient descent method [118] with these hyper-parameters: learning rate = 0.01, momentum = 0.9, drop-out rate = 0.1. Trainable weights were randomly initialized from a Gaussian distribution ($\mu = 0$, $\sigma = 0.01$) and updated with standard back-propagation.

6.4 Experiment

6.4.1 Annotated Datasets

Cardiac MRI data with gold-standard image quality annotations is available for circa 5,000 volunteers of the UKBB imaging resource. On the basis of experts' visual inspection, a simple three-grade quality score [21] is used for manual annotation: (1) optimal quality for diagnosis (4,361 sequences), (2) sub-optimal quality yet analysable (527 sequences), and (3) bad quality and diagnostically unusable (177 sequences). Since these data have full coverage of the heart from base to apex, all data with optimal quality was used to construct the ground-truth classes in our experiments. It is worth noting that data with full coverage of the heart does not mean the top or the bottom slice corresponds exactly to the base or apex.

To evaluate the performance of dataset adversarial learning, we use 598 MRI subjects obtained from the Cardiac Atlas Project (CAP) [65] (see Table. 6.1). The CAP is a web-accessible resource (http://www.cardiacatlas.org/), which provides a resource for cardiac image data sharing and atlas-based shape analysis for population studies. The datasets used in this study are part of two cohorts: MESA and DETERMINE. The imaging protocols included cine images acquired in short-axis planes from the base of the heart to the apex and in three long-axis planes. We manually checked one slice above and below the detected basal slice to make sure it is the right one and generate the annotations, ditto for apical slice checking.

For evaluating of models with PI (LAX patches) input, the LAX image information was extracted by collecting pixels values along the intersecting line between the 4-chamber view plane and corresponding short-axis plane over the cardiac cycle. We extracted 4 pixels above and below the two plane intersection. We embedded the constructed profile within a square image with zeros everywhere except the profile diagonal (see Fig. 6.2a bottom channel).

6.4.2 Data Augmentation and Resampling

Data Augmentation: To prevent over-fitting due to insufficient target data (DETERMINE and MESA), and to improve the detection rate of our algorithm, we employ data augmentation techniques to artificially enlarge the target datasets. For this purpose we chose a set of realistic rotations, scaling factors, and corresponding mirror images, and applied them to the MRI images. The set of rotations chosen were -45° and 45°, and the scaling factors 0.75 and 1.25. This increased the number of training samples by a factor of

eight. After data augmentation, we had 2400, and 2384 sequences for DETERMINE and MESA datasets, respectively.

Plane Pose Parameters: The ground-truth of the slice position d_b (the distance to basal slice), d_a (the distance to apical slice) and orientation parameterized with deflection angles θ in *xoy* plane and γ in z direction can be obtained from the 2D SAX images in the realistic 3D cardiac volumes. According to the Table. 6.1, the inter slice spacing value δd is constant in each cardiac volume, thus we can represent d_a and d_b by multiple δd. The orientation is chose based on θ and γ, which are calculated as the following process: 1) Standard Cardiac Vector: we choose the LV apex point C_L^A, the LV blood pool central points of each middle slice, and the middle point of mitral valve C_L^B in basal slice, using the coordinate of these points to fit the standard cardiac vector $\widehat{O}_L^S$ [117]; 2) Normal Vector: the normal vector $\widehat{O}_L^A$ is perpendicular to the SAX image plane; 3) Deflection Angles θ and γ: θ is the angle between the x axis and the projection of the deflection vector at *xoy* plane, γ is the angle between $\widehat{O}_L^S$ and $\widehat{O}_L^A$, which is calculated as $\gamma = \angle\langle \hat{O}_L^S, \hat{O}_L^A \rangle$.

Training set: We choose the UKBB image data and construct the multi-label (y^{t_d}, y^{t_θ}, y^{t_γ}) with a set of realistic distance and orientation values in the images for training our model. For each cardiac volume, we normalize the distance between basal slice and apical slice as unit 1, setting the distance label of basal slices as 0 and the distance label of apical slice as 1, then the label of the rest slices can be synthesised using $y_i^{t_d} = (i - Z_b)/(Z_a - Z_b)$ (Z_a and Z_b stand for the apical and basal slice numbers respectively in the sequence Z). Here, the normalized y^{t_d} can not only represent d_b, but also represent d_a. Based on analysis of the in-plane orientation angles distribution for 5,000 subjects for which manual segmentations are available (and therefore y^{t_θ}, y^{t_γ} can be computed), we found that y^{t_θ} ranges at the median value of 132.8° with standard deviation 8.0°, y^{t_γ} ranges at the median value of 7.1° with standard deviation 3.9°. The set of orientation labels were chosen from these realistic distributions and make all used labels balanced by data augmentation.

Testing set: During testing, we extract every slice from top to bottom for each volume and apply them into the the DARN* model. Our model ouput gives the slice position in millimeters and two angles (θ and γ) in degrees. There is no LAX patches (PI) in this phase.

6.4.3 Evaluation Metrics

We verify the effectiveness of our DARN* model through two groups of experiments. In the first experiment, the DIAL model is evaluated using a binary classification model instead of the MLMT* to detect the basal/apical slice. In the second experiment, we evaluate the MLMT* model using the fully fledged DARN*.

To evaluate the detection process, we measure classification accuracy, we use the following established classification metrics: Precision = $TP/(TP+FP)$, Sensitivity = $TP/(TP+FN)$, Error Rate = $(FP+FN)/N$, where TP, FP, and FN are numbers of true positive, false positive, and false negative samples, respectively, and N represents the number of subjects in the test set.

To evaluate the regression error between the predicted pose and the ground truth, we adopte Mean Absolute Error (MAE) and Cumulative Score (CS), which are two widely used performance methods, to evaluate the different models in our experiments. MAE computes the absolute costs between the exact and the predicted slice position or orientation (the lower the better): $MAE = \sum_{i=1}^{M} e_i/M$, where $e_i = \left|\widehat{l_i} - l_i\right|$ is the absolute cost of misclassifying true label l_i to $\widehat{l_i}$, and M is the total amount of testing samples. CS indicates the percentage of data correctly classified in the range of $(l_i - L, l_i + L)$, a neighbor range of the exact position or orientation label l_i (the larger the better): $CS(L) = \sum_{i=1}^{M} [e_i \leqslant L]/M$, where $[\cdot]$ is the truth-test operator and L is the parameter representing the tolerance range. Also, we used paired t-test to demonstrate the statistical significance of our empirical comparison if our DARN* significantly outperforms other methods.

6.4.4 CMR Slice Pose Estimation Results

To fully evaluate the effectiveness of the proposed method in different datasets, we conduct comprehensive comparison our approach with several state-of-the-art (related) approaches for cross datasets slice position estimation:

- **MC+CNN**: Metric Classification with CNN [271]
- **MC+CNN***: MC+CNN with PI [122]
- **MCDA+CNN**: MC-CNN with dataset invariance [70]
- **MCDA+CNN***: MCDA+CNN with PI
- **MR+CNN**: the metric regression CNN in [172]
- **DARN-DA**: DARN* without DA and PI (MLMT)
- **DARN*-DA**: DARN* without DA
- **DARN**: DARN* without PI
- **DARN*-MLMT**: DARN* without MLMT network [274]

- **DARN***: Fully fledged DARN* method

In particular, MC+CNN can be cast as a fundamental baseline only considering the deep classification neural network and MR+CNN can be cast as a fundamental baseline only considering the deep regression neural network. MCDA+CNN is the most relevant and state-of-the art cross-dataset image classification approach. For clarity, ablation study [101] is adopted to validate the effectiveness of our DARN* method by removing parts of the fully fledged model. We consider three special cases of the proposed method by excluding dataset invariance (DARN*-DA) or excluding privileged information (DARN*-PI) or excluding MLMT regression neural network (DARN*-MLMT) [274] for proving that each of the added term is useful for more accurate pose estimation.

1) Results Analysis for Basal and Apical slice Detection: To evaluate the performance of Dataset Invariance (DI) and Privileged Information (PI), we propose a baseline method dealing with the object detection problem, which only keeps the End-to-End CNN learning part and drops the part of transforming framework, *i.e.*, it casts the basal/apical slice detection problem as a metric classification problem, and addresses it with/without dataset invariance (adaptation vs non-adaptation) and Privileged Information, by transferring object classifiers from the UKBB to MESA and DETERMINE. For clarity, we compared the Metric Classification with CNN (MC+CNN) in [271], MC+CNN* [122] and the GTSRB architecture in [70] (MCDA+CNN) with our MCDA+CNN*. Table 5.2 shows the detection accuracy of the adaptation and non-adaptation for traditional CNN and CNN with PI. For both target datasets, the best improvements are the result of combining both of these features (DI plus PI). For MESA the detection accuracy was increased by 64%, and for DETERMINE best improvements are of 44% (right-most column). All the experiments are significantly different at $p < 0.05$.

2) Results Analysis for Slice Pose Estimation: We propose another model, which only keeps the End-to-End CNN learning part with dataset-invariance and PI, and drops the part of transforming framework, *i.e.*, it casts the metric classification module instead of a regression module, which transforms slice pose estimation, including the distance and orientation, into multi-label multi-task (MLMT) regression problem. To find out what factor gives more contributions to the final improvement of performance and validate that our regularization terms are beneficial, we also compared our proposed DARN* with DARN-DA, DARN*-DA, DARN*-PI and show a set of results in Table. 6.2 for position estimation results and Table. 6.3 for orientation regression estimation results.

Table 6.2 shows the average estimation errors of slice distance by the MAE metric in millimeter (*mm*). Even without using the PI-input channels, our dataset invariance framework is able to reduce the slice distance estimation error to less than half the average slice

Table 6.2: Regression error comparison between adaptation and non-adaptation methods, each with single (SAX)- and PI inputs for cardiac SAX slice position estimation in terms of MAE (Mean ± standard deviation)(mm). Best results are highlighted in bold. All experiments trained with UKBB data.

Dataset	No dataset adaptation			With dataset adaptation		
	MR+CNN [172]	DARN-DA	DARN*-DA	DARN	DARN*-MLMT	DARN* (Ours)
UKBB	5.43±1.4	4.11±1.6	**3.12**±1.1	4.98±1.9	3.86±1.9	3.41±1.9
MESA	8.21±1.6	7.94±2.0	6.53±1.7	4.97±1.0	3.91±1.1	**3.68**±1.1
DETERMINE	7.42±1.3	6.47±1.9	5.96±1.3	4.77±1.6	4.27±1.0	**4.05**±1.0

spacing found on our test datasets, *i.e.*, $< 5mm$. With PI in the training process we reduced the MAE to 4.27 and 4.05mm on average for slice position estimation. Table 6.3 shows the MAE of slice orientation estimation by regression θ and γ in degree ($^\circ$). Even without using the PI-input channels, our dataset invariance framework is able to get smaller estimation errors, *i.e.*, $\Delta\theta < 7^\circ$ and $\Delta\gamma < 4^\circ$. With PI input we reduced the estimation errors of $\Delta\theta$ and $\Delta\gamma$ to 5.24° and 3.45° on average for each volume in DETERMINE.

The comparison in terms of CS of the five state-of-the-art methods and the different combination of our algorithm. Clearly, DARN* outperforms all others across the entire range of L_d, L_θ and L_γ from 1 to 10. Specifically for DETERMINE, DARN* can reach the accuracy of 84.7% for $L_d = 8mm$, 80.1% for $L_\theta = 9^\circ$ and 77.9% for $L_\gamma = 6^\circ$. The other fact we notice is that four regression-based methods reach a higher accuracy for $L_d = 10mm$, $L_\theta = 10^\circ$ or $L_\gamma = 10^\circ$ than the others. All the experiments are significantly different at $p <$0.05.

6.5 Discussion

Automatic LVRV coverage estimation of CMR volumes is important in high-throughput image analysis of population imaging. Importantly, acquisition of thousands of suboptimal CMR images for later image analysis could be avoided if such quality assessment is performed on-line and a system provides immediate feedback to technical staff at the point of acquiring new images. Incomplete LVRV coverage and incorrect cardiac orientation influences the accuracy of ventricle anatomical and functional parameters of clinical interest. Manual annotation of cardiac pose is laborious, time-consuming and error prone in current clinical routine. To automate this labor-intensive task, we propose an efficient and robust framework for automatic across dataset estimation of CMR slice pose. Our framework has two main tasks: in the first task, we train the DIAL model that computes the common

Table 6.3: Comparison between adaptation and non-adaptation methods, each with single view (SAX) and PI inputs for cardiac SAX slice orientation estimation in terms of MAE (Mean $\pm$ standard deviation)($\Delta\theta$ and $\Delta\gamma$ indicate the MAE of the deflection angles in degree ($^\circ$). Best results are highlighted in bold.

Dataset	No dataset adaptation ($\Delta\theta/\Delta\gamma$)			With dataset adaptation ($\Delta\theta/\Delta\gamma$)		
	MR+CNN [172]	DARN-DA	DARN*-DA	DARN	DARN*-MLMT	DARN* (Ours)
UKBB	5.94±1.4/3.68±1.6	5.42±1.6/3.37±1.9	**5.25**±1.1/**3.24**±1.7	6.18±1.6/3.64±1.9	5.64±1.6/3.61±1.9	5.68±1.6/3.43±1.9
MESA	7.32±2.6/5.02±1.9	6.78±2.0/4.83±2.4	6.47±1.7/4.54±1.4	6.31±2.0/3.92±2.4	6.12±1.6/3.86±1.9	**5.96**±1.7/**3.74**±1.4
DETERMINE	7.14±2.1/5.11±2.3	6.63±1.9/4.81±2.0	6.32±1.3/4.17±1.4	6.27±1.6/3.81±1.3	5.82±1.6/3.69±1.9	**5.24**±1.0/**3.45**±1.3

representation for different datasets. It also learn image features from different MRI viewing planes across CMRI datasets to learn the appearance for the prediction of the different slice planes pose. The second task robustly estimates the slice pose based on the learned common representation using the MLMT* for the target cardiac volumes, and can also assist radiologists by automatically labeling potentially incomplete volumes to mark them for closer inspection. Extensive experimental results illustrate the effectiveness and efficiency of our method: its performance is superior to other methods with obvious advantages.

In any automatic image quality assessment system for population imaging, accuracy and robustness are key design criteria. These methods must work without many false positives or false negatives for basal/apical slice detection and the MAE of the slice distance/orientation should be small, and have to cope with considerable image quality variation. Most machine learning methods can achieve a high recognition accuracy by training and testing on single dataset. However, this can be prohibitive with different databases or when retraining is required as new data comes available. In this work, we used a very large dataset comprised of over 5,000 individually annotated cardiac MRI scans of the same number of subjects and each with 50 time points, which is 50-fold the 100 cases used in our previous work [271]. However, when transfer our well trained model to other CMRI datasets, deep learning methods without dataset invariance cannot achieve a good performance. We had to design an efficient network learning common representation across datasets. Considering there is no label information in target datasets, we also need to learn the discriminative information from source dataset and transfer them to our target datasets. Adversarial learning has been amongst the most promising solutions for reducing the difference between the training and test domain distributions and improve generalization performance. However, most adversarial learning works have been focused on image generative tasks, and little effort has been devoted to minimize an approximate domain discrepancy distance. We propose a novel adversarial learning to detect and localize the basal/apical slices across datasets, which incorporate the PI (cross-view information) into the training phrase. Then a MLMT regression network is trained to estimate the slice position and orientation. Specifically, our proposed DIAL and PI learning strategy can achieve a high accuracy rate of nearly 87%/90% for MBS/MAS detection by training on UKBB and testing on MESA, which is better than the CNN methods without dataset adaptation. Meanwhile, with the MLMT network, DARN* can decrease the MAE by around 6% compared with DARN*-MLMT [274] approaches for basal slice distance estimation in MESA dataset.

Our proposed automatic plane pose assessment framework for cardiac cine MRI has great potential to improve the robustness of later population image parsing. One could

imagine an approach whereby image analysis is adaptive to image quality and where different models are used depending on whether the volume under analysis incomplete ventricle coverage or incorrect cardiac orientation. In our architecture, we focus on learning common representation across datasets and develop a MLMT regression network to detect those that best discriminate slice positions and orientations. The advantages of an adversarial learning based representation for vision tasks are manifold: they can be composed to create common representations among various datasets of specificity; they are generalizable, as they can be learned once across datasets and then applied to recognize new objects or categories with no further training. One of our future work is to investigate the possibility of quantifying the ventricle coverage, not specific for slice position and orientation estimation, so we can predict the percentage of ventricle coverage directly. The difficulty of calculating the percentage of ventricle coverage lies in the different shape of contiguous ventricle slices, which makes the training of the volumes classifier a non-trivial task. Another future work is to extend deep-learning method for synthesizing the missed slices, *i.e.*, synthesizing the basal/or apical slice if a cardiac sample without them and the missing slices acquired from different positions. This is a limitation of our proposed framework, which can only estimate the slice positions and orientations. One possible way to achieve deep learning approach for image synthesis would be to apply generative adversarial network and synthesis the missed slices using adversarial approach on UKBB.

6.6 Conclusion

In this chapter, we have proposed a Dataset Adversarial Regression Network with Privileged Information (DARN*) framework capable of learning a common image representation, and using it to detect and estimate CMR slice pose, we achieve this by: first, using a DIAL model to fit the joint distribution over the images from different datasets with a minimax game. Second, extending the DIAL model to handle PI input scenarios thereby obtaining better results for slice pose estimation in cardiac MRI. And third, by introducing a MLMT regression network to predict the slice poses. We evaluated our framework on three large datasets UKBB, MESA and DETERMINE and found that our approach significantly outperforms state-of-the-art non-dataset-adaptive and non-PI methods. Finally, Our DARN* framework can be easily generalized to any anatomical structure or image modality.

Chapter 7

Quality-Aware Generative Adversarial Nets for Cross-Dataset Cardiac Cine MRI Synthesis

This chapter is based on:

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *MSIGAN: Missing Slice Imputation for Cardiac Cine MRI via Conditional Generative Adversarial Net*, accepted by MICCAI 2019. (In Press)

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *SPSGAN: Standard Plane Synthesis in Cardiac Cine MRI via Unsupervised Cycle-Consistent Adversarial Networks*, accepted by MICCAI 2019. (In Press)

Authors' contributions: L.Z., M.P. and A.F.F. conceived and designed the study; S.E.P., S.N. and S.P. provided support on clinical aspects and he also provided the UK Biobank data resource to be used for training and testing; L.Z. designed the method, performed data analysis and wrote the manuscript. All authors read and approved the manuscript.

7.1 Introduction

In this chapter, we present a new problem in medical image analysis, namely missing data imputation, and its specific problem in cardiac MR image analysis tasks, such as missing slice imputation (MSI) and standard cardiac plane synthesis (SPS). In the real world applications of big data processing, many tasks suffer a common drawback, missing or unknown data (incomplete feature vector). Specifically, in medical diagnosis, some examinations cannot be carried out because hospitals lack necessary medical equipment or some medical examinations are not suitable for some patients. Many studies have proposed different approaches to solve the incomplete data problems. One of these approaches is to impute or estimate the missing data, then, process the data using the edited set, *i.e.*, complete data portion and incomplete patterns with imputed values.

Accurate ventricular volume measurements depend on the complete heart coverage and correct cardiac orientation in CMR sequences that provide most immediate indicators of normal/abnormal cardiac function. However, incomplete heart coverage, especially missing basal/or apical slice, and the slice in CMR sequences with incorrect cardiac orientation (ICO) are substantial problems that are not sufficiently addressed in current clinical research and have an important impact on volume calculation. In this chapter, we propose two new deep architectures, one is called missing slice imputation generative adversarial network (MSIGAN), to learn the features of cardiac SAX slices cross different positions, and take the features as conditional variables to effectively infer the missing slices in the query volumes. Another one is called standard plane synthesis in cardiac cine MRI via unsupervised cycle-consistent adversarial networks (SPSGAN), which given a SAX slice with ICO, automatically generates images under correct orientation. In MSIGAN, the slices are first mapped to latent vectors with position features through a regression net, and then the latent vector with desired position is projected to the condition on slice intensity through a generator net. The latent vector preserved with the slice features (*i.e.*, intensity) and the desired position condition control the generation vs. regression. Two adversarial networks are imposed on the regressor and generator, respectively, forcing to generate more realistic slices. In SPSGAN, we address this challenge by dividing the problem into two subtasks. First, we consider using a bidirectional generator that maps the initially rendered image back to an image with input cardiac orientation, which can be directly compared to the input image without requiring any GT images. Second, to generate high perceptual quality images, we propose a novel loss function that incorporates intensity and orientation terms.

7.2 MSIGAN: Missing Slice Imputation for Cardiac Cine MRI via Conditional Generative Adversarial Net

Cardiac MRI can not only reflect anatomic information of the heart but also provide physiological information associated with cardiovascular diseases. The EF and CO of the both ventricle, defined by the difference between basal and apical slices, are the most commonly used clinical diagnostic parameters for cardiac myocardium function and cardiac volume calculation. Most published studies have addressed this classification problem by assuming that a complete data set with all features available for all samples. In practice, this assumption is not valid because some tests may be missed due to high measurement costs or lack of patient consent [236]. However, full set of features are required for every sample in the training and testing datasets when training the discriminative classifiers (*i.e.*, SVM).

A common strategy to deal with the incomplete data is to delete them from the study cohort [256] [236]. However, removing data not only reduces statistical ability, but also raises ethical concerns because the subject data obtained are not yet used. Recently, some data imputation based methods are proposed to deal with this problem, such as using data's mean or model-based missing data estimation [75]. If the missing mechanism is random, the missing variable can be imputed by the marginal distribution of the observed data using the maximum likelihood estimation (MLE) [56]. The stochastic regression imputation method can make better use of the information provided by the data to solve the collinearity problem caused by the high correlation of predicted variables [208]. when the data missingness is non-random, the missing variable cannot be predicted only from the available variables in the database, and there is no general method of handling missing data properly [75]. The performance of imputation approaches is ideally assessed by both the feature error and the classification accuracy on the imputed features.

In this section, we adapt the developed generative adversarial net (GAN) to generate the missing slices after applying the quality control (QC). We propose a MSI-based generative adversarial network (MSIGAN) model to infer missing slice features from multi-position images input. After inference, the feature of the desired position and the slice intensity feature are concatenated for further generating real images in certain positions. The main contributions of the MSIGAN are highlighted as follows:

(1) A novel deep MSIGAN architecture is proposed for generating missing SAX slices for cardiac MRI across different positions. A regression net learns intrinsic features of the input volume firstly. Conditioned on these features as well as a pre-computed feature of

the expected position, generator and discriminator aim to generate real images of the same volume in expected position.

(2) Given the feature of the slice in expected position, we design a conditional generative network to infer an image matching with the missing slice of the input cardiac volume. The adversarial training mechanism and auxiliary slice position regressor is combined to achieve effective feature generation.

(3) This is the first paper exploit the deep learning method, especially GAN, for missing slice imputation in cardiac MRI, which is an important step after QC and before quantitative medical image analysis. It can be learned once and then applied to synthesize the missing slice for incomplete heart coverage without any further training.

7.2.1 Methodology

Problem Formulation: The overall target of MSI for cardiac MRI is similar with the missing data imputation problem in the field of data mining [166]. Given a query cardiac MR volume, a regression list of slice positions in the gallery set is desired, processing images synthesis for the query volumes slices where it is missed. For each input a 3D cardiac image $\mathbf{X}$, we aim to map its feature to a representation $\mathbf{f}$ and synthesize the missing slice $\hat{\mathbf{x}}$ by the following function:

$$\hat{\mathbf{x}} = \Gamma(series(\{\mathbf{f}_n\}_{n=1}^{N})) = \Gamma(series(\mathscr{R}(\mathbf{X}) \cdot \{\Upsilon_n\}_{n=1}^{N})). \tag{7.1}$$

Where the operator $\mathscr{R}(\cdot)$ is to extract the features (*i.e.*, intensity) of the input image $\mathbf{X}$. $\{\Upsilon_n\}_{n=1}^{N}$ is obtained by the regression model to identify the slice position features, like the distance to basal/or apical slice of the inferred slice. N is the number of slices. Moreover, the operator $\Gamma(\cdot)$ denotes the transformation from the concatenated features to the inferred slice features in the cardiac volume. Finally, we can synthesize the missed slice in a certain position. Therefore, the most significant factor to achieve effective synthesis is how to design and optimize the $\mathscr{R}(\cdot)$, Υ and $\Gamma(\cdot)$.

We formulate the image synthesis for MSI problem with the following three steps: first, Given an 3D cardiac volume $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ and the corresponding slice position label $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, a regression net $\mathscr{R}(\cdot)$ aims to learn the cardiac intensity feature $\mathbf{f}_{int}$ and the slice position maps Υ. Second, exploiting the feature maps for different slice positions and the intensities as conditions, we aim to generate desired slice features by $\Gamma(\cdot)$ with an adversarial training architecture. A generative net takes the intrinsic slice features (*i.e.*, intensity) and random vectors and position feature in desired position as inputs to synthesize the missed cardiac cine MRI. Third, a discriminative net distinguishes the generated

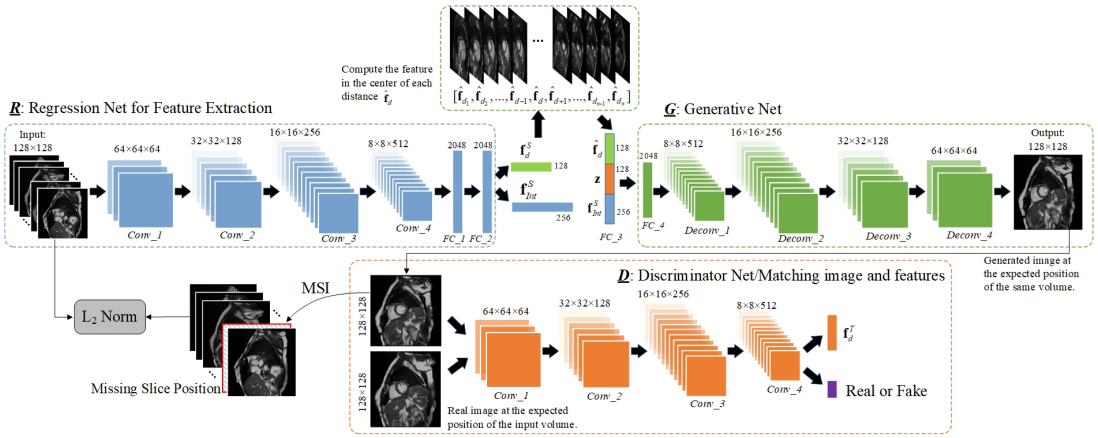


Figure 7.1: Structure of the proposed MSIGAN network for cardiac MSI. The regressor R maps each slice of the input volume to a vector containing intensity and position features. Moreover, the central point feature of each position cluster over the whole training set can be obtained and used for generator G. Concatenating the intensity feature and the random noise to the inferred position cluster center feature, the new latent vector FC_3 is fed to G. Both the R and G are updated based on the L_2 loss between the original and synthetic volumes. The discriminative net D forces the output slice to be realistic and plausible for a given position label.

samples from the real images, and simultaneously tries to match the inferred slices with correct features and positions. The network architecture is illustrated in Figure 7.1. The synthesized slices can be directly adopted for imputing the missed slice in the target CMR volumes.

Cardiac feature learning and slice position estimation: To generate CMR slices in SAX view, a deep regression network R aims to learn CMR image features including slice position $\mathbf{f}_d$ and intensity $\mathbf{f}_{Int}$. Formally, $\mathbf{X}_S$ denotes the input cardiac image stack with full ventricular coverage. The trunk architecture of the regression net consists of 4 convolutional layers (kernel size = 5, padding = 2 and stride = 2) and 2 fully-connected layers. The Leaky-ReLU is set after each layer. Leaky-ReLU includes a very small slope for negative value inputs. This mitigates against dead neuros, as the derivative is always non-zero, allowing gradient based learning to occur (however slow). Then, we configure two layers for learning the 256-dimensional $\mathbf{f}_{Int}$ with the intrinsic intensity features and 128-dimensional $\mathbf{f}_d$ by inferred slice position regression separately, since we expect slice position information weakened in $\mathbf{f}_{Int}$, but strengthened in $\mathbf{f}_d$. During training the regression net, the loss function can be fast and well converged. Thus, we can easily learn each position's feature cluster from all the training data by k-means clustering, and compute the feature in the center of each cluster, $\mathbf{f}_{dc}$, as a condition to generate slices in the missed position.

Conditional Cardiac GAN: Instead of generating real images by normal GANs, our model aims to transform the features from CMR volumes with full ventricular coverage into the query CMR volumes, which miss slices in certain positions, by a generative model. The conditional generator is defined as G: $\mathbb{R}^F \times \mathbb{R}^Z \times \mathbb{R}^T \rightarrow \mathbb{R}^S$, where F is the dimension of intrinsic cardiac intensity, Z is for random noise, T is the dimension of inferred slice position and S is for cardiac slice. Besides, the discriminator is denoted as D: $\mathbb{R}^S \rightarrow \{0,1\} \times \prod l_i$, where $i = \{1 : \mathbf{f}_{Int}, 2 : \mathbf{f}_{dc}\}$. l_i denotes the range of each label. The optimization of the G and D can be reformulated as:

$$\mathscr{L}_D = E_{\mathbf{x} \sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})] - \sum_{i=1}^{2} \|l_i - D(\mathbf{x}))\|_2^2 \tag{7.2}$$

$$\mathscr{L}_G = E_{A;B;C}[\log\Big(1 - D(G(\mathbf{f}_{dc}^T, \mathbf{z}, \mathbf{f}_{Int}^S))\Big)], \tag{7.3}$$

where

$$\begin{aligned} A &\rightarrow \mathbf{f}_{dc}^T \sim p_{data}(\mathbf{f}_{dc}^T), \\ B &\rightarrow \mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}), \\ C &\rightarrow \{\mathbf{f}_{Int}^S\} \sim p_{data}(\mathbf{f}_{Int}^S). \end{aligned}$$

The input of the generator G is the concatenation of the $\mathbf{f}_{Int}^S, \mathbf{f}_{dc}^T$ and a random noise prior $\mathbf{z} \sim \mathscr{N}(0,1)$. $\mathbf{f}_{Int}^S$ can be regarded as intrinsic intensity features from the *original* slices, while $\mathbf{f}_{dc}^T$ are the *desired* position feature in the center of the cluster. A fully-connected layer is set for better fusing the three vectors and then four deconvolutional layers are adopted for generating synthesized slice samples. The hyper-parameter settings of the generative net are reverse to that of the regression net R.

The discriminator D takes the generated samples and the real images in the target CMR volume as inputs. The main structure of D has the similar structure in regression net. To match the inferred slices with the same intensity features and correct slice position in the query volumes, we add a fully-connected layer and simultaneously optimize the whole discriminative net by slice position regression. The position label for the synthetic slice is same with the expect position label in the query volume. Batch normalization and ReLU are adopted for all the layers in the discriminator as well. Meanwhile, to ensure the output slice sharing the intensity with the input image (during training), the input image and output image are expected to be similar as expressed in Equation. (7.4), where $L(\cdot)$ denotes L_2 norm.

$$\mathscr{L}_{L2N} = L(\mathbf{x}, G(R(\mathbf{x}))) \tag{7.4}$$

Optimization: The training scheme for MSIGAN consists of three steps. In the first step, R is trained using a deep regression net for slice feature learning. Then, the computed

different slice position features are obtained. In the second step, G is fed by the learned real position features from different cardiac volumes, which is fused with the intensity features $\mathbf{f}_{Int}^{S}$ and the random noise $\mathbf{z}$. Four deconvolutional layers are adopted for G to generate synthesized slice samples. Both the regressor and the generator are updated based on the L_2 loss between the input and output volumes to ensure they are similar. In the following step, the discriminative net D employs a general fully convolutional network to distinguish the real images from the generated ones. Rather than maximizing the output of the discriminator for generated data, the objective of feature matching [204] is employed to optimize G to match the statistics of features in an intermediate layer of D. The objective function is defined in the following equation:

$$\begin{aligned}\mathscr{L}_{MSI} = &\min_{G}\max_{D,R} E(\log\Big(1-D(G(\{\mathbf{f}_{Int}^{S},\mathbf{z},\mathbf{f}_{dc}^{T}\}))\Big)) + \sum_{i=1}^{2}\|l_i - D(\mathbf{x}))\|_2^2 \\ &+ \Big\|E(D_k(\{\mathbf{f}_{Int}^{S},\mathbf{f}_{d}^{S}\})) - E(D_k(G(\{\mathbf{f}_{Int}^{S},\mathbf{z},\mathbf{f}_{dc}^{T}\})))\Big\|_2^2 - L(\mathbf{x},G(R(\mathbf{x})))\end{aligned} \tag{7.5}$$

where k means the k^{th} layer in D (k = 4 in our setting). Moreover, D is trained with slice position regression to better match generated position features with input volume's identities. We apply one more *conv* layer to output the final position features. For all the *conv* layers in G and D, we adopt Leaky-ReLU activation and batch normalization. The conditioned G and D nets can be optimized by $\mathscr{L}_{MSI}$ to infer the missing features from query input volumes.

7.2.2 Experiments and Analysis

Materials and Position Label Generation. Quality-scored CMR data is available for circa 5,000 volunteers of the UKBB imaging resource. Following visual inspection, manual annotation for SAX images was carried out with a simple 3-grade quality score [21]. 4,280 sequences correspond to quality score 1 for both ventricles, these had full coverage of the heart from base to apex and were the source datasets to construct the ground-truth distance label for our experiments. Note that having full coverage should not be confused with having the top/bottom slices corresponding exactly to base/apex [274].

The slice position labels are generated from the realistic distances to apex point and base point. To obtain the apex point, we take last 2 apical 2D manual delineations and fit a spline curve to extrapolate location of the apex. Then measure distance to apex for all image slices; to obtain the base point, we use LA manual delineations on the 4CH LA image view to define the center of the MV as the base. Then measure distance from this point to all image slices. To make each slice label represents the distance to apex and base

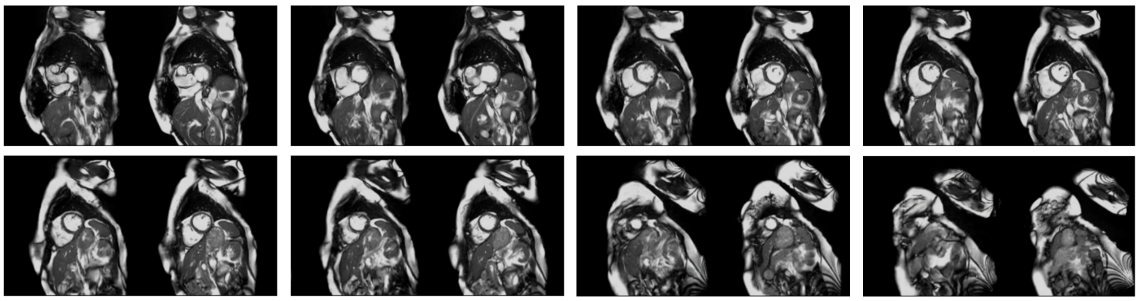


Figure 7.2: Example of synthesized images (*left*) generated by MSIGAN, compared to the GTs (*right*).

simultaneously, we normalize the distance from base to apex as unit 1 for all cases (base as 0 and apex as 1), and label the middle slices with values equally increased from 0 to 1. For the slices above base and below apex, we also use the equal interval to label them.

Experimental Settings. We performed two groups of experiments in this work. In the first experiment, we aim to evaluate the quality of the images generated by MSIGAN. The averaged peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index are used to measure the image quality of those ground-truth (GT) and synthetic MR images. In the second group, we evaluate the imputed cardiac volumes with corresponding GT on both tasks of LV segmentation and measurement of cardiac function based on blood volumes. Four parameters are used for performance evaluation, including two commonly used indexes of the cardiac function derived from such volumes viz. SV and EF, and similarly report the differences between the real and imputed coverages.

Performance of Image Synthesis Model. To evaluate the quality of the images generated by MSIGAN, we first train the MSIGAN model using 3,280 complete subjects from the 4,280 cases with quality score 1 in UKBB, and test our model on the rest 1,000 subjects. We take the 1,000 testing subjects for which the GT slices are available and randomly remove the slices to generate incomplete volumes, and using our MSIGAN to synthesize the missed slices. Several typical images with real and synthetic slices are shown in Figure. 7.2. We can observe that our synthetic slices look very similar to their corresponding real images. Also, the mean and standard deviation of PSNR and SSIM values of synthetic slice are listed in Table 7.1. Results indicate that MSIGAN performs much better than other methods based on the other metrics of PSNR and SSIM for missing slice synthesis. These results imply that our trained MSIGAN model is reasonable, and the synthetic cardiac MRI scans have acceptable image quality (in terms of PSNR and SSIM).

Results of Cardiac Functional Parameters Calculation. To assess the impact of synthetic images in real applications, such as measurement of cardiac function based on blood

Table 7.1: Quantitative results for missing cardiac MRI synthesis based on PSNR and SSIM. Higher values indicate better performance. Values in bracket represent standard deviation across volumes. Absolute highest performing results seen in bold.

	Mean	GMM [256]	SCGAN [272]	MSIGAN (ours)
PSNR	20.49±5.21	22.17±3.75	17.49±3.46	**24.49±3.69**
SSIM	0.547±0.21	0.686±0.24	0.512±1.71	**0.703±0.11**

Table 7.2: Effect of incomplete cardiac coverage (MBS) on the ED, ES, SV and EF. Values are shown as Mean ± standard deviations.

	Ground Truth	Missing Basal Slice (MBS)	Effect(%)	Synthetic Image	Effect(%)
LVEDV(ml)	155.8±35.6	136.1±33.4	-12.6%	151.7±33.7	**-2.6%**
LVESV(ml)	66.8±21.2	53.0±19.0	-20.7%	61.3±22.3	**-8.2%**
LVSV(ml)	89.1±19.8	83.1±19.7	-6.7%	90.4±18.7	**+1.5%**
LVEF(%)	57.1±0.06	61.0±0.06	+6.8%	59.6±0.06	**+4.4%**

volumes, we design an experiment where incomplete coverage is simulated and volume differences between ground-truth, synthetic volumes and incomplete volumes are measured. The experimental results achieved by seven different cardiac parameters using LV segmentation method in [238] are reported in Table 7.2. For this experiment, we compute blood pool volumes at the ED and ES phases, and from these, we obtain SV and EF. Then, the average volumes and indexes are computed across the sample, comparing the ground-truth, synthetic volumes and incomplete volumes. Table 7.2 shows that MBS reduces ED and ES volumes by an average of 12% and 20%, respectively. In contrast, the synthetic values are much closer than the GT values, with 2.6% and 8.2% reduction in volumes at ED and ES phases. These results clearly demonstrate that the synthetic images generated by our MSIGAN model are useful in clinical application. Significant differences between each methods (p <0.05) are indicated respectively.

7.2.3 Conclusion

In this section, we proposed a deep MSIGAN to implement missing cardiac cine MRI generation and contribute to the missing data imputation neglected by the medical imaging community. The MSI adopts a slice position regression model and the adversarial training architecture to impute those missing slices based on their corresponding distances to base and apex, considering the relationship between neighbour slices scanned for the same subject. Extensive experimental results showed that our model could both achieve satisfactory performance on missing slice generation and imputation compared to some baselines. Our

method are reasonable to practical applications. Currently, only the complete images are used for learning the segmentation models. Using these synthetic slice data could further augment the training samples for improvement, which will be our future work.

7.3 SPSGAN: Standard Plane Synthesis in Cardiac Cine MRI via Unsupervised Cycle-Consistent Adversarial Networks

Cardiac MRI can not only reflect anatomic information of the heart but also provide physiological information associated with cardiovascular diseases. This requires the careful selection of consistent orientation of short-axis (SAX) image planes with respect to the cardiac ventricles such as the basal slice (BS) and apical slice (AS) plane that contain key anatomical structures [273]. If the plane orientation deviates significantly from expected values, local image structure may change enough to cause subsequent image feature-based algorithms to fail in localizing key features required for further morphological and functional analysis. However, it is challenging and time-consuming even for experienced MRI scanner to manually navigate the machine to find the correct standard plane. The task is highly operator-dependent and requires a great amount of expertise. With the advent of cardiac MRI, 2D slice in SAX view can be acquired quickly with little training. But the problem of locating diagnostically required standard planes for biometric measurements remains. There is a strong need to develop automatic methods for 2D standard plane generation from existing 2D slices to improve clinical workflow efficiency.

Image Generation is a hot topic which has achieved great success in many vision tasks, such as text-to-image generation [191] and image style transformation [140]. Generative Adversarial Networks (GANs) [81] adopt a convolutional network (discriminator) to a deconvolutional network (generator) in order to improve the performance of the generator in learning a realistic data distribution while trying to confuse the discriminator. It has shown impressive results in rendering new realistic images. Conditional Generative Adversarial Networks (cGANs) [162] is more advanced in image generation and more suitable for image translation tasks. It is developed by adding the input condition vector, which can include vast amount of information, to the generator. Medical image synthesis is currently an emerging area of interest for application of the latest image generation techniques mentioned above. Nie et al. [170] proposed a context-aware GANs by adding an image gradient difference term to the loss function of the generator, with the aim of retaining the sharpness of the generated images. Dar et al. [46] utilized CycleGAN and pix2pix technique in generating T1-weighted MR contrast from T2-weighted MR contrast or vice versa.

Inspired by above ideas, we propose a GAN framework using fully unsupervised approach, which given a SAX slice with ICO, automatically generates images under standard orientation. To train this model using unlabeled data (*i.e.*, our training data consists of the

query images and the images with correct orientation of different cardiac volumes), we propose a Cycle-GANs based architecture that combines a novel loss function that transfers the plane orientation and generate new images of high perceptual quality [76]. The main contributions of the SPSGANs are highlighted as follows:

(1) A novel deep SPSGAN architecture is proposed for generating SAX images with correct plane orientation. To achieve this, we devised a novel loss function computed over the images used in a Cycle-GANs for orientation transfer.

(2) Unlike the traditional Cycle-GANs, we proposed an unsupervised strategy that is trained in the absence of paired examples for imageto-image translation.

(3) This is the first paper exploit the deep learning method, especially GAN, for orientation based cardiac slice generation, which is an important step after QC and before quantitative medical image analysis.

7.3.1 Methodology

Problem Formulation: In order to produce realistic standard orientation transformations of the input slice while retaining the intensity appearance, we use a single SAX slice as input and train a GAN model using an unsupervised approach. Formally, we seek to learn the mapping $(\mathbf{x}_t^i, \mathbf{f}_{\theta_o}, \mathbf{f}_{\gamma_o}) \rightarrow \mathbf{x}_o^i$ between an image $\mathbf{x}_t^i \in \mathbb{R}^{H\times W\times Z}$ with incorrect plane orientation $< \theta_t, \gamma_t >$ and the image $\mathbf{x}_o^i \in \mathbb{R}^{H\times W\times Z}$ with the standard plane orientation $< \theta_o, \gamma_o >$ and same cardiac identity. Orientations are represented by $< \theta, \gamma >$, where θ indicates the deflection angle in *xoy* plane and γ indicates the deflection angle in the z direction of the 3D coordinate, respectively. The subscript o and t denote as the *standard (correct)* and *transformed (incorrect)* orientations, respectively. The model is trained using an unsupervised approach with training samples $\{\mathbf{x}_o^i, \mathbf{x}_t^j\}_{i,j=1}^N$, which do not include the GT image $\mathbf{x}_o^j$.

SPS Unsupervised Cardiac GAN (SPSGAN): Figure. 7.3 shows the structure of our SPSGAN model. It consists of five main modules: (1) The learned real orientation features $[\mathbf{f}_\theta, \mathbf{f}_\gamma]$ of images from different cardiac volumes, which are concatenated with the features in the generator for better generating an image with desired orientations. (2) A generator $G(\mathbf{x}\,|(\mathbf{f}_\theta, \mathbf{f}_\gamma))$ that maps one given slice under an incorrect orientation to an output slice under the correct (standard) orientation with the same cardiac identity. Note that G is used twice in our network, first to map the input image $\mathbf{x}_{tr}^i \rightarrow \mathbf{x}_{og}^i$ and then render the latter back to the initial orientation $\mathbf{x}_{og}^i \rightarrow \hat{\mathbf{x}}_{tg}^i$; (3) A regressor R responsible of estimating the slice orientation of a given image. Note that R is different from the pre-trained regression net for feature extraction in (1); (4) A discriminator D that tries to discriminate the generated and real images; (5) A loss function that aims to preserve the cardiac intensity by computing

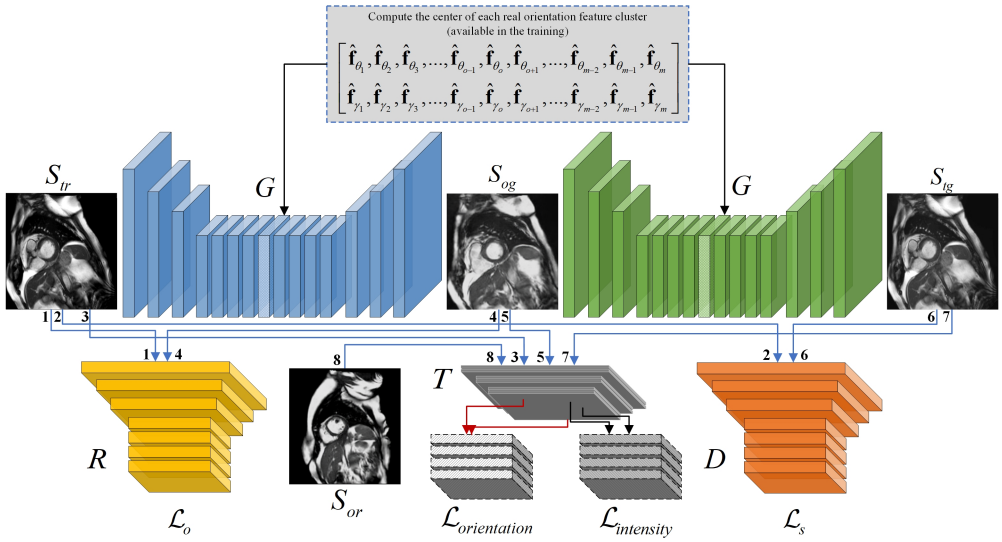


Figure 7.3: The structure of our SPSGAN to generate standard plane of cardiac MRI in SAX view. Our model consists of five main components: a generator G, a discriminator D, an orientation regressor R, the transfer net T and the pretrained orientation features. Neither GT image is considered.

without GT. To address this challenge, we propose a novel loss function that enforces intensity content similarity of $\mathbf{x}_{tr}^i$ and $\hat{\mathbf{x}}_{tg}^i$, and orientation similarity between $\mathbf{x}_{og}^i$ and $\mathbf{x}_{or}^j$. In the following, we describe in detail each of the five modules.

Real Orientation Feature Embedding: The orientations of all volumes are linearly distributed and categorized with two parameters $< \theta, \gamma >$. During training the regression net, the loss of orientation regression can be fast and well converged. Thus, we can easily learn the each orientation feature cluster from all the training data by k-means clustering, and compute the feature in the center of the cluster, as a condition to generate images in desired orientation. The feature of each orientation in an image $\mathbf{x}_t^i \in \mathbb{R}^{H\times W\times Z}$ is represented as a probability density map $\mathbf{f}$ computed over the entire image domain as:

$$\mathbf{f} = \Gamma(series(\{\mathbf{x}_n\}_{n=1}^N)). \tag{7.6}$$

Where the operator $\Gamma(\cdot)$ is to extract the feature of the input image $\mathbf{x}$. N is the defined number of slices.

Generator: Given an input image $\mathbf{x}$, the generator $G(\mathbf{x}\,|(\mathbf{f}_\theta, \mathbf{f}_\gamma))$ aims to render the input slice of an incorrect orientation in a standard orientation with $< \theta_o, \gamma_o >$. To condition the generator with the orientation features we consider the concatenation $(\mathbf{x}, \mathbf{f}_\theta, \mathbf{f}_\gamma) \in \mathbb{R}^{H\times W\times Z}$ and feed this into a feedforward network, which generates output images of the same size as $\mathbf{x}$. To achieve an impressive results for the image-to-image translation, we adopt the variation of the network from [105] to construct the generator.

Image Discriminator: We adopt the PathchGAN [102] network as the discriminator $D(\mathbf{x})$, which maps from the input image $\mathbf{x}$ to a matrix $Y_s \in \mathbb{R}^{26\times 26}$, and the discriminator tries to classify if each 26×26 patch in an image is real or fake. Since a smaller PatchGAN can generate high perceptual quality images with fewer parameters and less time [102], we run the discriminator across the image with convolutional manner and average all responses to provide the final output D.

Orientation Regressor: The D distinguishes the generated samples from the real images, and simultaneously we use an orientation regressor R tries to regress the inferred slice with correct orientations. R is implemented with the ResNet architecture in [280].

Optimization: We have three terms to be optimized for the full loss function. An generative adversarial loss that enforces the distribution of the generated image to be similar to that of the training image. An orientation regression loss that enforces the orientation of the generated images to be similar to the standard orientation. The transfer loss that preserve the cardiac identity between the generated and the input images. Next, we will describe each of these terms.

Generative Adversarial Loss: To optimize the parameters of generator G and learn the distribution of the training data, we perform a standard *minmax* strategy game between the generator and the image discriminator D. The generator and discriminator are jointly trained with the objective function $\mathscr{L}_s(G,D,\mathbf{x},\mathbf{f}_\theta,\mathbf{f}_\gamma)$ where D tries to maximize the probability of correctly classifying original and rendered images while G tries to foul the discriminator.

$$\mathscr{L}_s(G,D,\mathbf{x},\mathbf{f}_\theta,\mathbf{f}_\gamma) = E[\log D(\mathbf{x})] + E[\log\big(1 - D(G(\mathbf{x}\,\big|\,(\mathbf{f}_\theta,\mathbf{f}_\gamma)))\big)] \tag{7.7}$$

Orientation Regression Loss: The generator G not only reduces the generative adversarial loss, but also must reduce the error produced by the orientation regressor R. In this way, while learning to produce realistic samples, G also learns how to generate images consistent with the standard orientation $< \theta, \gamma >$. This loss is defined by:

$$\mathscr{L}_o(G,R,\mathbf{x},\mathbf{f}_\theta,\mathbf{f}_\gamma) = \big\|R(G(\mathbf{x}\,\big|\,(\mathbf{f}_\theta,\mathbf{f}_\gamma))) - <\mathbf{f}_\theta,\mathbf{f}_\gamma>\big\|_2^2 \tag{7.8}$$

Transfer Loss: With the two previously defined losses $\mathscr{L}_s$ and $\mathscr{L}_o$, G is enforced to generate realistic cardiac slices with correct orientation. However, in the absence of GT supervision, there is no constraint to ensure the appearance identity. We derive inspiration from the previously introduced content-style loss to maintain high perception quality in image style transfer [76]. The loss mainly consists of two parts, one retains intensity similarity and the other transfers orientation similarity. Inspired by this idea, we define two sub-losses to maintain the identity between the input slice $\mathbf{x}_{tr}^i$ and the rendered slice $\mathbf{x}_{og}^i$.

For the intensity term, we define that G should be able to render-back the initial slice $\mathbf{x}^i_{tr}$ given the generated slice $\mathbf{x}^i_{og}$ and the original orientation features $< \mathbf{f}_{\theta_t}, \mathbf{f}_{\gamma_t} >$, that is $\hat{\mathbf{x}}^i_{tg} \approx \mathbf{x}^i_{tr}$, where $\hat{\mathbf{x}}^i_{tg} = G(G(\mathbf{x}^i_{tr} \,|(\mathbf{f}_{\theta_o}, \mathbf{f}_{\gamma_o})) \,|(\mathbf{f}_{\theta_t}, \mathbf{f}_{\gamma_t}))$. Nevertheless, even when using PatchGAN based discriminators, directly comparing $\mathbf{x}^i_{tr}$ and $\hat{\mathbf{x}}^i_{tg}$ at a pixel level would struggle to handle highfrequency details leading to overly-smoothed images. Instead, we compare them based on their intensity content. Formally, we define the intensity loss to be:

$$\mathscr{L}_{intensity} = \left\| T_l(\mathbf{x}^i_{tr}) - T_l(\mathbf{x}^i_{tg}) \right\|_2^2 \tag{7.9}$$

where $T_l(\cdot)$ represents the feature representation at the l^{th} layer of the network.

In order to transfer the standard orientation information of the real slice into the rendered one, we take over the spatial extent of the feature maps to design the feature space for capturing texture information. As previous work [76] implement this by computing the Gram matrix $\mathbf{M}^l \in \mathbb{R}^{U \times U}$, where $\mathbf{M}^l$ is the inner product between the vectorised feature maps of $\mathbf{x}^i_{og}$. The orientation loss is then computed as the mean square error between visible pairs of Gram matrices of the same joint in both images $\mathbf{x}^i_{og}$ and $\mathbf{x}^j_{or}$:

$$\mathscr{L}_{orientation} = \frac{1}{L} \sum_{l=0}^{L} \left(\frac{\mathbf{M}^{i,l}_{og} - \mathbf{M}^{j,l}_{or}}{UV} \right)^2 \tag{7.10}$$

where $\mathbf{M}^{i,l}_{og}$ and $\mathbf{M}^{j,l}_{or}$ are the orientation representation in the layer l of the generated image and the real image with standard orientation, respectively. In layer l, there is U_l feature maps each of size V_l, where V_l is the height times the width of the feature map. Finally, we define the transfer loss as the weighted sum of the intensity and orientation losses:

$$\mathscr{L}_{TS} = \mathscr{L}_{content}(T, \mathbf{x}^i_{tr}, \hat{\mathbf{x}}^i_{tg}) + \lambda \mathscr{L}_{orientation}(T, \mathbf{x}^i_{tr}, \mathbf{x}^i_{og}, \mathbf{x}^j_{or}) \tag{7.11}$$

where he parameter λ controls the relative importance or the two components.

Full Loss: We take the full loss as a linear combination of all previous loss terms:

$$\mathscr{L}_{SPS} = \arg\min_{G} \max_{D,R,T} \{ \alpha \mathscr{L}_s(G, D, \mathbf{x}, \mathbf{f}_\theta, \mathbf{f}_\gamma) + \beta \mathscr{L}_o(G, R, \mathbf{x}, \mathbf{f}_\theta, \mathbf{f}_\gamma) + \mathscr{L}_{TS} \} \tag{7.12}$$

where α and β are the weighting factors for image adversarial and orientation regression loss, respectively.

7.4 Experiments and Analysis

Materials and Evaluation Metrics. There are 5,000 CMR subjects available in the UKBB imaging resource and each volumetric sequence contains about 50 cardiac phases.

Based on analysis of the in-plane orientation angles distribution for the 5,000 subjects for which manual segmentations are available (and therefore θ, γ can be computed), we found that θ ranges at the median value of 132.8° with standard deviation 8.0°, γ ranges at the median value of 7.1° with standard deviation 3.9°. Among them, there are 302 cases under standard cardiac orientations ($\theta = 135^\circ, \gamma = 0^\circ$). The set of orientation labels were chosen from these realistic distributions and trained in a regression net to obtain the real orientation features.

Since our SPSGAN model is trained using the unsupervised approach, we need to generate the slices with correct orientation as GT for the slices under incorrect orientations to evaluate the synthetic images. The GT images are resampled from the interpolated 3D cardiac volumes by Paraview[1]. The resampled slices are chosen with correct orientations and the same position (*i.e.*, the distance to base and apex) compared with the original images.

Experimental Settings. We verify the effectiveness of our unsupervised SPSGAN model through two groups of experiments. In the first experiment, the synthetic slice is evaluated against the GT using rotation angles ($\delta\theta$ viz. $\delta\gamma$) between the planes. Image similarity of the planes is also measured using peak signal-to-noise ratio (PSNR) and structural similarity (SSIM). In the second group, we evaluate the synthetic slices with corresponding GT on both tasks of LV segmentation and measurement of cardiac function based on blood volumes. Four parameters are used for performance evaluation, including two commonly used indexes of the cardiac function derived from such volumes viz. SV and EF, and similarly report the differences between the real and imputed coverage.

Performance of Image Synthesis Model. We train the SPSGAN model using the 302 subjects with correct orientation and same number of cases with incorrect orientations in UKBB, and test the model on another 100 subjects with incorrect orientations and the corresponding resampled GT. Training images are only associated to the original slices with correct and incorrect orientations. No GT images are considered during training. Several typical images with real and synthetic slices are shown in Fig. 7.4. We can observe that our synthetic images show a slight different with their corresponding original images, but similar with their corresponding GT images. This is because the local image structure in planes with different orientations will change. Also, the orientation angles, SSIM and PSNR between synthetic, original and GT slices are shown in Fig. 7.4. These results imply that our trained SPSGAN model is reasonable, and the synthetic CMR images have acceptable representation for the standard planes.

Results of Cardiac Functional Parameters Calculation. To assess the impact of synthetic images in real applications, such as measurement of cardiac function based on blood

[1]https://www.paraview.org/

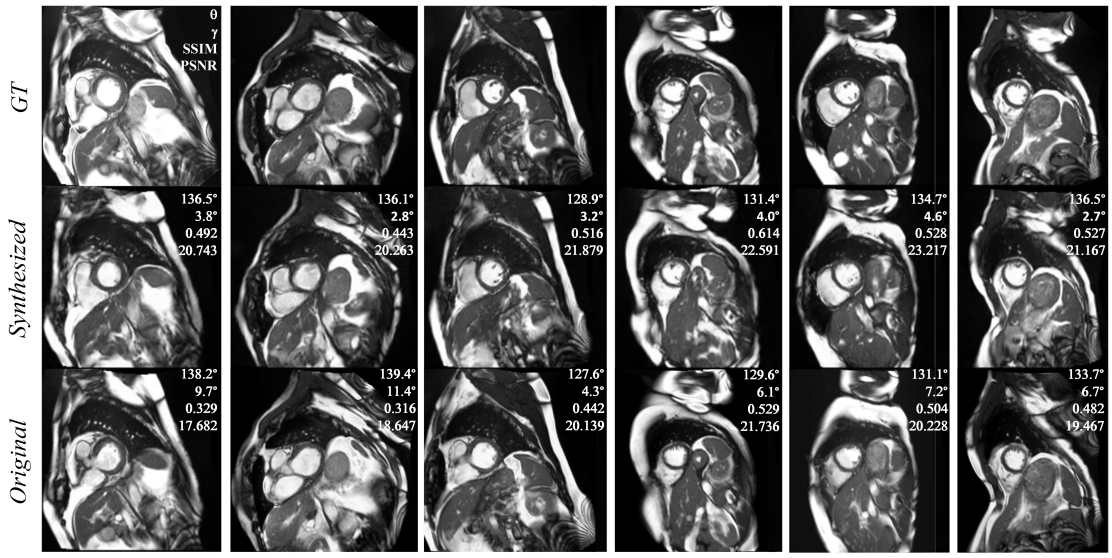


Figure 7.4: Example of synthesized images generated by SPSGAN and the corresponding original images with orientation angles, PSNR and SSIM values, compared to the GTs.

Table 7.3: Effect of ICO on the ED, ES, SV and EF. Values are shown as Mean ± standard deviations.

	Ground Truth	Synthetic Image	Effect(%)	ICO Image	Effect(%)
LVEDV(ml)	159.6±32.7	151.5±34.9	-5.1%	142.9±31.5	-10.5%
LVESV(ml)	72.4±23.1	68.3±20.3	-5.7%	64.3±22.4	-11.2%
LVSV(ml)	87.2±17.6	83.2±18.4	-4.6%	78.6±17.9	-9.9%
LVEF(%)	54.6±0.08	54.9±0.09	+0.5%	55.0±0.08	+0.7%

volumes, volume differences between GT, synthetic volumes and volumes with incorrect cardiac orientation are measured in this experiment. The experimental results achieved by seven different cardiac parameters using LV segmentation method in [238] are reported in Table 7.3. For this experiment, we compute blood pool volumes at the ED and ES phases, and from these, we obtain SV and EF. Then, the average volumes and indexes are computed across the sample, comparing the GT, synthetic volumes and incomplete volumes. Table 7.3 shows that the incorrect plane orientation reduces ED and ES volumes by an average of 10% and 11%, respectively. In contrast, the synthetic images are much closer than the GT values, with 5.1% and 5.7% reduction in volumes at ED and ES phases. These results clearly demonstrate that the synthetic images generated by our SPSGAN model are useful in clinical application.

7.5 Conclusion

We have presented a novel approach for generating cardiac cine MRI under standard plane orientation using a GAN model that can be trained using a fully unsupervised approach. Finding the correct standard plane is highly operator-dependent and requires a great amount of expertise. To tackle this challenge, we proposed an new framework that aims at retaining the cardiac orientation and intensity of the original image instead of training the data by optimizing a loss function that only depends on the input image and the GT. Extensive experimental results showed that our model could both achieve satisfactory performance on standard cardiac slice generation compared to some baselines. Our method are reasonable to practical applications. In the future, we plan to further exploit our approach in other datasets (not only of UKBB) in different modalities for which supervision is not possible.

Chapter 8

Summary and Future Work

Here, we give a brief overview of the scope of the thesis and the related research problems. In addition, we summarize briefly the proposed and achieved technical contributions. We also discuss the current limitations of the proposed algorithms and propose possible directions for future research. Finally, we conclude the thesis with a general outlook on possible cardiac image computing research.

8.1 Summary and Achievement

Biomedical imaging has become an indispensable and increasingly important component of disease diagnosis and treatment. MRI is a well-established imaging modality that is widely used in clinical examination, diagnosis, treatment, and decision-making; however, in practice, MRI is susceptible to a variety of artifacts that reduce image quality, possibly leading to inefficient and/or inaccurate diagnoses. Sources of artifacts in MRI include nonideal hardware characteristics, intrinsic tissue properties and possible changes in them during scanning, assumptions underlying data acquisition and image reconstruction processes, and poor selection of scanning parameters [139]. To minimize or eliminate artifacts, automatic methods for such repetitive quality assurance tasks provide the required consistency and reliability.

CMR imaging is an increasingly common technique for clinical diagnostic imaging of the heart. For population imaging studies, CMR remains the modality of choice and provides all-in-one, noninvasive access to cardiac anatomy and function [182]. The quantification of LV anatomy and function from large population imaging studies or from patient cohorts from large clinical trials requires automatic image quality evaluation and image analysis tools. Basic criteria for cardiac image quality include ventricle coverage and the detection of missing apical and basal CMR slices [118]. However, full automation and reliable ventricle coverage assessment for CMR images face certain challenges.

- Detecting complete/incomplete ventricle coverage, especially missing basal slice and missing apical slice, as well as finding the position of the basal and apical slices.
- Regression for SAX slices to find the distance and orientation of each slice.
- Synthesis of the missing slices to recover image quality based on detection and regression results.

This thesis has focused on the adverse effects of such challenges imposed by CMR images on common automated analysis tasks, such as image classification, regression, and image synthesis. The experimental results demonstrate that the proposed frameworks can yield more reliable solutions by overcoming the challenges observed in cardiac imaging.

8.1.1 Image Feature Learning for LV Coverage Assessment

CMR images are playing an increasing role in the diagnostic imaging of CVDs. Full LV coverage, from base to apex, is a basic criterion for CMR image quality and is required for accurate cardiac volume measurement and functional assessments. Incomplete LV coverage is identified through visual inspection, which is time consuming and typically performed retrospectively in the assessment of large imaging cohorts. In Chapter 3, we discussed using a 2D CNN constructed on single-slice images; the images were then processed sequentially. However, this solution ignores the contextual information contained across slices, which provides inferior performance compared with 3D analysis. In Chapter 4, we proposed an automatic method to determine LV coverage from CMR images using an FD3D CNN. In contrast to our previous method that uses 2D CNNs, this approach attempts to learn feature representations to achieve reliable classification results even with small amount of training data or a limited number of iterations. Our FD3D CNN utilizes 3D convolution kernels and exploits spatial contextual information in volumetric data by building an end-to-end architecture of CNNs. The proposed FD3D CNN uses the FD criterion in the fully-connected layer to make the features discriminative and insensitive to geometric structural variations.

At the end of the Chapter 4, we described extensive experiments performed to validate the proposed method on more than 5,000 independent volumetric CMR scans from a UKBB study. The result demonstrated low error rates for missing basal/apical slice detection (4.9%/4.6%). To the best of our knowledge, this is the first study to tackle the problem of automatic detection of missing apical and basal slices in CMR imaging in an evaluation using a very extensive and challenging population imaging dataset. The proposed method may also be adapted for LV coverage assessment of other types of CMR image data. Future

studies can benefit from the FD based classification method to investigate this problem in more depth.

8.1.2 Adversarial Cross-dataset Feature Learning

Cardiac functional parameters, such as EF and CO of both ventricles, are the most immediate indicators of normal or abnormal cardiac function. To compute these parameters, accurate measurement of ventricular volumes at ED and ES is required. Accurate volume measurements depend on the correct identification of basal and apical slices on CMR sequences that provide full LV and RV coverage. In Chapter 5, we proposed a CNN-based AL approach that detects and localizes the basal/apical slices in an image volume independently of image acquisition parameters, such as imaging device, magnetic field strength, and variations in protocol execution. The proposed model is trained on multiple cohorts of different provenance and learns image features from different MRI viewing planes to learn the appearance and predict the position of the basal and apical planes. To the best of our knowledge, this is the first study to address fully-automatic detection and position regression of basal/apical slices in CMR volumes in a dataset-invariant manner. This was achieved by maximizing CNN's ability to regress the position of basal and apical slices within a single dataset while minimizing the classifier's ability to discriminate image features between different data sources. The results demonstrate superior performance compared with state-of-the-art methods.

8.1.3 Image Feature Learning for Slice Pose Estimation

CMR imaging is the standard imaging technique used to evaluate morphology and functionality of the heart. After acquisition, automatic techniques can be used to extract volumetric information and derive clinical indexes that place a subject within the predetermined population ranges of normality. Accurate volume measurements depend on the correct identification of ventricle pose, especially the slice positions and orientations, in CMR sequences that provide full LV and RV coverage. In Chapter 6, we proposed a CNN-based AL approach that regresses the pose of CMR slices in an image volume independently of image-acquisition parameters, such as imaging device, magnetic field strength, and variations in protocol execution. We incorporate additional information, such as cross-view image information into the training phrase, and refer to this information as PI. The proposed model is trained on multiple cohorts of different provenance and unified by a novel PI loss with different MRI viewing planes to learn the appearance and to correctly orient the short-axis view planes of the heart. To the best of our knowledge, this is the first study to tackle

fully-automatic detection and pose estimation of biventricular slices in CMR volumes in a dataset-invariant manner. We achieve this by maximizing CNN's ability to regress the position and orientation of short-axis view planes within a single dataset while minimizing the classifier's ability to discriminate image features between different data sources. The results show superior performance compared with the existing state-of-the-art methods.

8.1.4 Image Feature Learning for Missing Data Imputation

Accurate ventricular volume measurements depend on complete heart coverage and correct cardiac orientation in CMR sequences that provide the most immediate indicators of normal/abnormal cardiac function. However, incomplete heart coverage, especially missing basal/or apical slices, and slices in CMR sequences with ICO are substantial problems that affect the volume, but are not sufficiently addressed in current clinical research. In this thesis, we propose two new deep architectures. One is called MSIGAN, which is used to learn the features of cardiac SAX slices across different positions and to consider the features as conditional variables to effectively infer missing slices in query volumes. The other one is called SPSGAN, which provided with a SAX slice with ICO, automatically generates images under correct orientation. In a MSIGAN, slices are first mapped to latent vectors with position features through a regression net and then the latent vector with the desired position is projected to the slice manifold, conditional on slice intensity through a generator net. The latent vector preserved with the slice features (*i.e.*, intensity) and the desired position condition control generation versus regression. Two adversarial networks are imposed on the regressor and generator, forcing generation of more realistic slices. In a SPSGAN, we address this challenge by dividing the problem into two principal subtasks. First, we consider a bidirectional generator that maps back the initially rendered image to the original orientation, hence being directly comparable to the input image without the need to resort to any training image. Second, we devise a novel loss function that incorporates intensity and orientation terms, and aims at producing high perceptual quality images.

8.2 Limitations and Future Work

In this section, the most significant limitations of the presented image analysis methods are discussed together with suggestions to tackle these limitations in future research. It is important to note that the technical drawbacks are not limited to the items described in the following sections.

8.2.1 Transfer Learning

The most important potential of DL methods lies in their ability to extract a series of discriminative features from multi-layer neural networks. As mentioned previously, a CNN is a supervised learning model, and training CNNs from scratch requires extensive memory and computing resources, otherwise the training process will be extraordinarily time consuming. DL requires a large number of labeled training data, and manual labeling by medical experts is both time consuming and very expensive. In addition, in some cases, such as tumors, few images are available. Training deep models often becomes very complex due to over fitting and convergence problems. It is often necessary to repeatedly adjust the learning parameters of the network. To overcome this challenge, future research could focus on pretraining deep models in a supervised manner using natural images or could employ the transfer learning method to enable the use of datasets from different medical fields. For example, we can pre-train the deep model on ImageNet dataset and fine-tuning the parameters, then use it for medical image classification problems.

8.2.2 Unsupervised/Weakly-Supervised Learning

The existing literature indicates that most of the advanced DL methods, particularly CNN-based frameworks, involve supervised learning approaches. Previous studies have focused on pretrained CNNs or on using CNNs as a feature extractor, which can be easily downloaded and directly applied to medical image analysis. In medical image analysis, end-to-end training of CNNs has become the preferred method. However, applying DL methods in medical data analysis is problematic because obtaining sufficient annotated data for supervised learning is a major challenge [254] [200]. Given that labeled training data is limited, developing methods that can use non-labeled images will be the focus of future work. In addition, future studies should focus on developing weakly supervised learning methods that combine the advantages of supervised and unsupervised learning. For example, we can use the limited lablled data in UKBB to develop the weakly-supervised model and then test it on large number of unlabelled data.

8.2.3 Data Harmonization

Typically, datasets collected by researchers and clinicians, particularly those working in different locations, are often recorded using different formats and protocols. Thus, without previous reorganization and preprocessing, such datasets cannot be used directly by various computing technologies, such as the algorithms developed in this thesis. Nevertheless, consistently and completely achieving this remains a technical challenge. For example, in

Chapters 6 and 7, we discussed automatically cleaning, computing and reorganizing the data to include slice distance and orientation information in the UKBB such that the image input can be read by the slice pose regression pipeline. Even while working with two widely used public databases, MESA and DETERMINE, several challenges must be overcome. Therefore, to utilize computational techniques being developed to analyze imaging data and conduct effective large-scale data analysis, the development of automated methods to clean and organize data is an important issue.

8.2.4 Metadata Generation

Another important challenge is a method to automatically generate and use appropriate metadata to effectively describe how the data is recorded. Metadata is a powerful tool to annotate and exploit image-related information for clinical and research purposes. Metadata can be used to organize and archive images and to retrieve images and associated data from archives. Metadata organizing systems can minimize human burden and will enable standardization of formats, which will facilitate subsequent data analysis. For example, text-based reports of medical experts and electronic medical records contain rich clinical information that could be used to supplement labeled image data. It is expected that, combined with RNNs and CNNs, the natural image subtitle generation methods used in the computer vision field will be applied to medical image analysis in the near future.

In this thesis, we have reported the development of techniques to extract intensity-level anatomical information from large-scale cardiac image studies. However, this information only describes a single aspect of health and disease. The environment, lifestyle, biochemistry, and genetics can provide additional important complimentary information. As population studies become established, an important research direction will be to integrate different data sources and scales of biology and physiology to provide a systemic description of health and disease.

8.2.5 Knowledge Extraction and Interpretation

Once image data is suitably accessed by algorithms (supported by data harmonization algorithms and standard metadata formats), processed to extract anatomical information, and integrated with other sources of data, the next natural challenge for researchers will be to interpret the data for clinical purposes. Analysis of patient data for knowledge extraction will help deliver better healthcare with regard to disease diagnosis and prognosis, as well as treatment stratification. For example, in case of cardiac research, tools to extract more advanced and more complex phenotypes of cardiac health and disease are required.

Many studies have proven the importance of LV EF in predicting prognosis [44]. Other studies have established the relationship between LV wall thickness and higher mortality [53] [206]. Big data analysis will facilitate the study of new cardiac function indicators and multi-source biomarkers, which, in turn, will enable early identification of patients at risk for cardiovascular events, which will contribute to reducing mortality and morbidity rates in developed countries.

In summary, big data analytics is an emerging trend that is creating new opportunities, as well as challenges. In future, we will see rapid and extensive implementation and use of big data analysis by healthcare organizations and healthcare industry. To that end, the challenges highlighted above must be addressed in a holistic manner. As big data analytics becomes more widespread, it will attract more attention regarding the importance of establishing standards and continually improving related tools and technologies.

In this context, proper analysis of large amounts of medical image data is an important step toward developing methods to process and extract the information that can be derived from this data. Thus, the quality assessment and analysis tools that have been presented in this thesis represent an important step toward achieving effective big data analytics of all medical related information. Big data analytics and its application in healthcare are at a nascent stage of development; however, advances in tools can accelerate their maturation.

List of Publications

Book Chapter

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Image Quality Assessment for Population Cardiac MRI*, Deep Learning and Convolutional Neural Networks for Medical Image Computing. Edited by Le Lu, Xiaosong Wang, Gustavo Carneiro and Lin Yang. In Press.

Journal Papers

- **Le Zhang**, Ali Gooya, Marco Pereañez, Bo Dong, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Automatic Assessment of Full Left Ventricular Coverage in Cardiac Cine Magnetic Resonance Imaging with Fisher Discriminative 3D CNN*, IEEE Transactions on Biomedical Engineering, 66(7), 1975-1986 (July 2019).

- Rahman Attar, Marco Pereañez, Ali Gooya, Xènia Albà, **Le Zhang**, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Alejandro F. Frangi, *Cardiac Population Image Quantification: Analysis of 20K Subjects in the UK Biobank*, Medical Image Analysis, 56, 26-42 (August 2019).

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Automatic Plane Pose Estimation Across Cardiac Cine MRI Datasets via Deep Adversarial Ranking Nets with Privileged Information*, Submitted.

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Quality-Aware Generative Adversarial Nets for Cross-Dataset Cardiac Cine MRI Synthesis*, Preparing.

Conference Papers

- **Le Zhang**, Marco Pereañez, Christopher Bowles, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Missing Slice Imputation in Population CMR Imaging via Conditional Generative Adversarial Nets*, MICCAI 2019, Accepted (AR $< 30\%$).

- **Le Zhang**, Marco Pereañez, Christopher Bowles, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Unsupervised Standard Plane Synthesis in Population Cine MRI via Cycle-Consistent Adversarial Networks*, MICCAI 2019, Accepted (AR $< 30\%$).

- **Le Zhang**, Marco Pereañez, Stefan Piechnik, Stefan Neubauer, Steffen Petersen and Alejandro F. Frangi, *Multi-Input and Dataset-Invariant Adversarial Learning (MDAL) for Left and Right-Ventricular Coverage Estimation in Cardiac MRI*, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 481-489, Springer, Cham, 2018. (AR $< 30\%$)

- Rahman Attar, Marco Pereañez, Ali Gooya, Xènia Albà, **Le Zhang**, Stefan K. Piechnik, Stefan Neubauer, Steffen E. Petersen, Alejandro F. Frangi, *High Throughput Computation of Reference Ranges of Biventricular Cardiac Function on the UK Biobank Population Cohort*, MICCAI Statistical Atlases and Computational Modeling of the Heart (STACOM) Workshop, pp. 114-121, Springer, 2018.

- **Le Zhang**, Ali Gooya, and Alejandro F. Frangi, *Semi-supervised assessment of incomplete LV coverage in cardiac MRI using generative adversarial nets*, MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 61-68. Springer, Cham, 2017

- **Le Zhang**, Ali Gooya, Bo Dong, Rui Hua, Steffen E. Petersen, Pau Medrano-Gracia, and Alejandro F. Frangi, *Automated quality assessment of cardiac MR images using convolutional neural networks*, MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 138-145. Springer, Cham, 2016

All manuscripts consist of full-length, 8+ page papers that undergo double-blinded peer-review by 3-7 experts in the field, with highly competitive acceptance rates (AR), which are stated, where available. Top Conferences, such as MICCAI, have lower acceptance rates than many top journals.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. F. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. TensorFlow: Large-scale Machine Learning on Heterogeneous Distributed Systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] A. H. Abdi, C. Luong, T. Tsang, G. Allan, S. Nouranian, J. Jue, D. Hawley, S. Fleming, K. Gin, J. Swift, et al. Automatic Quality Assessment of Echocardiograms using Convolutional Neural Networks: Feasibility on the Apical Four-Chamber View. *IEEE Transactions on Medical Imaging*, 36(6):1221–1230, 2017.

[3] S. S. Aboutalib, A. A. Mohamed, W. A. Berg, M. L. Zuley, J. H. Sumkin, and S. D. Wu. Deep Learning to Distinguish Recalled but Benign Mammography images in Breast Cancer Screening. *Clinical Cancer Research*, 24(23):5902–5909, 2018.

[4] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: Deep Learning from Crowds for Mitosis Detection in Breast Cancer Histology Images. *IEEE Transactions on Medical Imaging*, 35(5):1313–1321, 2016.

[5] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou. Lung Pattern Classification for Interstitial Lung Diseases using a Deep Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 35(5):1207–1216, 2016.

[6] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez. Convolutional Neural Networks for Mammography Mass Lesion Classification. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 797–800. IEEE, 2015.

[7] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. G. Lopez. Representation Learning for Mammography Mass Lesion Classification with Convolutional Neural Networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, 2016.

[8] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[9] A. K. Attili, A. Schuster, E. Nagel, J.H.R.J. Reiber, and R.J. van der Geest. Quantification in Cardiac MRI: Advances in Image Acquisition and Processing. *International Journal of Cardiovascular Imaging*, 26:27–40, 2010.

[10] M. Avendi, A. Kheradvar, and H. Jafarkhani. A Combined Deep Learning and Deformable-Model Approach to Fully Automatic Segmentation of the Left Ventricle in Cardiac MRI. *Medical Image Analysis*, 30:108–119, 2016.

[11] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson. From Generic to Specific Deep Representations for Visual Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–45, 2015.

[12] W. J Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, et al. Human-Level CMR Image Analysis with Deep Fully Convolutional Networks. *arXiv preprint*, 2017.

[13] Y. Bar, I. Diamant, L. Wolf, and H. Greenspan. Deep Learning with Non-Medical Training Used for Chest Pathology Identification. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140V. International Society for Optics and Photonics, 2015.

[14] Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen, and H. Greenspan. Chest Pathology Detection using Deep Learning with Non-Medical Training. In *ISBI*, pages 294–297. Citeseer, 2015.

[15] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[16] O. Bernard, J. G. Bosch, B. Heyde, M. Alessandrini, D. Barbosa, S. Camarasu-Pop, F. Cervenansky, S. Valette, O. Mirea, M. Bernier, et al. Standardized Evaluation System for Left Ventricular Segmentation Algorithms in 3D Echocardiography. *IEEE Transactions on Medical Imaging*, 35(4):967–977, 2016.

[17] Kaggle.com. (2016). Data Second Annual Data Science Bowl. Data Science Bowl Cardiac Challenge Data. `https://www.kaggle.com/c/second-annual-data-science-bowl/data`. Accessed 17 Mar. 2016.

[18] T. Brosch, L. YW. Tang, Y. Yoo, D. KB. Li, A. Traboulsee, and R. Tam. Deep 3D Convolutional Encoder Networks with Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation. *IEEE Transactions on Medical Imaging*, 35(5):1229–1239, 2016.

[19] R. W. Brown, E. M. Haacke, Y. C. N. Cheng, M. R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley & Sons, 2014.

[20] Y. Cai, M. Landis, D. T. Laidley, A. Kornecki, A. Lum, and S. Li. Multi-Modal Vertebrae Recognition using Transformed Deep Convolution Network. *Computerized Medical Imaging and Graphics*, 51:11–19, 2016.

[21] V. Carapella, E. Jiménez-Ruiz, E. Lukaschuk, N. Aung, K. Fung, J. Paiva, M. Sanghvi, S. Neubauer, S. Petersen, I. Horrocks, et al. Towards the Semantic Enrichment of Free-text Annotation of Image Quality Assessment for UK Biobank Cardiac Cine MRI Scans. In *Deep Learning and Data Labeling for Medical Applications*, pages 238–248. Springer, 2016.

[22] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered Multiview Mammogram Analysis with Pre-trained Deep Learning Models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 652–660. Springer, 2015.

[23] G. Carneiro, J. C. Nascimento, and A. Freitas. The Segmentation of the Left Ventricle of the Heart from Ultrasound Data using Deep Learning Architectures and Derivative-based Search Methods. *IEEE Transactions on Image Processing*, 21(3):968–982, 2012.

[24] J. C. Carr, O. Simonetti, J. Bundy, D. Li, S. Pereles, and J. P. Finn. Cine MR Angiography of the Heart with Segmented True Fast Imaging with Steady-State Precession. *Radiology*, 219(3):828–834, 2001.

[25] E. Castillo and D. A. Bluemke. Cardiac MR Imaging. *Radiologic Clinics*, 41(1):17–28, 2003.

[26] E. Castillo, J. AC. Lima, and D. A. Bluemke. Regional Myocardial Function: Advances in MR Imaging and Analysis. *Radiographics*, 23(suppl_1):S127–S140, 2003.

[27] C. C. Chang and C. J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[28] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *arXiv preprint arXiv:1405.3531*, 2014.

[29] H. Chen, Q. Dou, X. Wang, J. Qin, J. CY. Cheng, and P. A. Heng. 3D Fully Convolutional Networks for Intervertebral Disc Localization and Segmentation. In *International Conference on Medical Imaging and Virtual Reality*, pages 375–382. Springer, 2016.

[30] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng. Standard Plane Localization in Fetal Ultrasound via Domain Transferred Deep Neural Networks. *IEEE Journal of Biomedical and Health Informatics*, 19(5):1627–1636, 2015.

[31] H. Chen, L. Wu, Q. Dou, J. Qin, S. Li, J. Z. Cheng, D. Ni, and P. A. Heng. Ultrasound Standard Plane Detection using A Composite Neural Network Framework. *IEEE Transactions on Cybernetics*, 47(6):1576–1586, 2017.

[32] H. Chen, L. Yu, Q. Dou, L. Shi, V. CT. Mok, and P. A. Heng. Automatic Detection of Cerebral Microbleeds via Deep Learning Based 3D Feature Representation. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 764–767. IEEE, 2015.

[33] S. X. Chen, C. J. Zhang, M. Dong, J. L. Le, and M. Rao. Using Ranking-CNN for Age Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5183–5192, 2017.

[34] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[35] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng. Dual Path Networks. In *Advances in Neural Information Processing Systems*, pages 4467–4475, 2017.

[36] G. Cheng, P. Zhou, and J. Han. RIFD-CNN: Rotation-invariant and Fisher Discriminative Convolutional Neural Networks for Object Detection. In *Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition*, pages 2884–2893, 2016.

[37] H. Childs, L. Ma, M. Ma, J. Clarke, M. Cocker, J. Green, O. Strohm, and M.G. Friedrich. Comparison of Long and Short Axis Quantification of Left Ventricular Volume Parameters by Cardiovascular Magnetic Resonance, with Ex-vivo Validation. *Journal of Cardiovascular Magnetic Resonance*, 13:40, 2011.

[38] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *International Conferenceon Medical Image Computing and Computer-Assisted Intervention*, pages 424–432. Springer, 2016.

[39] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. Th. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken. Automatic Classification of Pulmonary Peri-Fissural Nodules in Computed Tomography using An Ensemble of 2D Views and A Convolutional Neural Network Out-of-the-box. *Medical Image Analysis*, 26(1):195–202, 2015.

[40] D. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images. In *Advances in Neural Information Processing Systems*, pages 2843–2851, 2012.

[41] D. Cireşan, U. Meier, and J. Schmidhuber. Multi-Column Deep Neural Networks for Image Classification. *arXiv preprint arXiv:1202.2745*, 2012.

[42] D. C. Cireşan, A. Giusti, L. M. Gambardella, and J. Schmidhuber. Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013.

[43] S. Conolly, A. Macovski, J. Pauly, J. Schenck, K. K. Kwong, D. A. Chesler, X. P. Hu, W. Chen, M. Patel, and K. Ugurbil. Magnetic Resonance Imaging. In *Medical Devices and Systems*, pages 243–282. CRC Press, 2006.

[44] J. P. Curtis, S. I. Sokol, Y. Wang, S. S. Rathore, D. T. Ko, F. Jadbabaie, E. L. Portnay, S. J. Marshalko, M. J. Radford, and H. M. Krumholz. The Association of Left Ventricular Ejection Fraction, Mortality, and Cause of Death in Stable Outpatients with Heart Failure. *Journal of the American College of Cardiology*, 42(4):736–742, 2003.

[45] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable Convolutional Networks. *CoRR, abs/1703.06211*, 1(2):3, 2017.

[46] S. H. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur. Image Synthesis in Multi-Contrast MRI with Conditional Generative Adversarial Networks. *arXiv preprint arXiv:1802.01221*, 2018.

[47] M. de Bruijne. Machine Learning Approaches in Medical Image Analysis: From Detection to Diagnosis, 2016.

[48] B. D. de Vos, M. A. Viergever, P. A. de Jong, and I. Išgum. Automatic Slice Identification in 3D Medical Images with A ConvNet Regressor. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 161–169. Springer, 2016.

[49] B. D. de Vos, J. M. Wolterink, P. A. de Jong, T. Leiner, M. A. Viergever, and I. Igum. ConvNet-Based Localization of Anatomical Structures in 3D Medical Images. *IEEE Transactions on Medical Imaging*, 36(7):1470–1481, July 2017.

[50] B. D. de Vos, J. M. Wolterink, P. A. de Jong, M. A. Viergever, and I. Išgum. 2D Image Classification for 3D Anatomy Localization: Employing Deep Convolutional Neural Networks. In *Medical Imaging 2016: Image Processing*, volume 9784, page 97841Y. International Society for Optics and Photonics, 2016.

[51] S. Demyanov. ConvNet Library for Matlab [Online]. `https://github.com/sdemyanov/ConvNet`. Accessed 15 Oct. 2017.

[52] J. Deshmukh and U. Bhosle. Image Mining using Association Rule for Medical Image Dataset. *Procedia Computer Science*, 85:117–124, 2016.

[53] R. B. Devereux, K. Wachtell, E. Gerdts, K. Boman, M. S. Nieminen, V. Papademetriou, J. Rokkedal, K. Harris, P. Aurup, and B. Dahlöf. Prognostic Significance of Left Ventricular Mass Change During Treatment of Hypertension. *Jama*, 292(19):2350–2356, 2004.

[54] J. Dolz, N. Betrouni, M. Quidet, D. Kharroubi, H. A. Leroy, N. Reyns, L. Massoptier, and M. Vermandel. Stacking Denoising Auto-Encoders in A Deep Network to Segment the Brainstem on MRI in Brain Cancer Patients: A Clinical Study. *Computerized Medical Imaging and Graphics*, 52:8–18, 2016.

[55] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial Feature Learning. *arXiv preprint arXiv:1605.09782*, 2016.

[56] Y. Dong and C. Y. J. Peng. Principled Missing Data Methods for Researchers. *SpringerPlus*, 2(1):222, 2013.

[57] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng. Multilevel Contextual 3D CNNs for False Positive Reduction in Pulmonary Nodule Detection. *IEEE Transactions on Biomedical Engineering*, 64(7):1558–1567, 2017.

[58] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. CT. Mok, L. Shi, and P. A. Heng. Automatic Detection of Cerebral Microbleeds from MR Images via 3D Convolutional Neural Networks. *IEEE Transactions on Medical Imaging*, 35(5):1182–1195, 2016.

[59] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P. A. Heng. 3D Deeply Supervised Network for Automated Segmentation of Volumetric Medical Images. *Medical Image Analysis*, 41:40–54, 2017.

[60] D. Erhan, P. A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent. The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-training. In *Artificial Intelligence and Statistics*, pages 153–160, 2009.

[61] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature*, 542(7639):115, 2017.

[62] P. F. Ferreira, P. D. Gatehouse, R. H. Mohiaddin, and D. N. Firmin. Cardiovascular Magnetic Resonance Artefacts. *Journal of Cardiovascular Magnetic Resonance*, 15(1):41, 2013.

[63] S. D. Fihn, J. M. Gardin, J. Abrams, K. Berra, J. C. Blankenship, A. P. Dallas, P. S. Douglas, J. M. Foody, T. C. Gerber, A. L. Hinderliter, et al. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS Guideline for the Diagnosis and Management of Patients with Stable Ischemic Heart Disease: A Report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines, and the American College of Physicians, American Association for Thoracic Surgery, Preventive Cardiovascular Nurses Association, Society for Cardiovascular Angiography and Interventions, and Society of Thoracic Surgeons. *Journal of the American College of Cardiology*, 60(24):e44–e164, 2012.

[64] J. P. Finn, K. Nael, V. Deshpande, O. Ratib, and G. Laub. Cardiac MR Imaging: State of the Technology. *Radiology*, 241(2):338–354, 2006.

[65] C. G. Fonseca, M. Backhaus, D. A. Bluemke, R. D. Britten, J. D. Chung, B. R. Cowan, I. D. Dinov, J. P. Finn, P. J. Hunter, A. H. Kadish, et al. The Cardiac Atlas ProjectAn Imaging Database for Computational Modeling and Statistical Atlases of the Heart. *Bioinformatics*, 27(16):2288–2295, 2011.

[66] M. Foppa, B. B. Duncan, and L. E. Rohde. Echocardiography-Based Left Ventricular Mass Estimation. How Should We Define Hypertrophy? *Cardiovascular Ultrasound*, 3(1):17, 2005.

[67] D. F. Fouhey, A. Gupta, and A. Zisserman. 3D Shape Attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1516–1524, 2016.

[68] A. F. Frangi, W. J. Niessen, and M. A. Viergever. Three-Dimensional Modeling for Functional Analysis of Cardiac Images, A Review. *IEEE Transactions on Medical Imaging*, 20(1):2–5, 2001.

[69] Kilem L. G. Intrarater Reliability. In *Wiley Encyclopedia Clinical Trials*, pages 1–14. Hoboken, NJ, USA: Wiley, 2008.

[70] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, et al. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[71] M. Gao, U. Bagci, L. Lu, A. Wu, M. Buty, H. Shin, H. Roth, G. Z. Papadakis, A. Depeursinge, R. M. Summers, et al. Holistic Classification of CT Attenuation Patterns for Interstitial Lung Diseases via Deep Convolutional Neural Networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(1):1–6, 2018.

[72] X. Gao, W. Li, M. Loomes, and L. Wang. A Fused Deep Learning Architecture for Viewpoint Classification of Echocardiography. *Information Fusion*, 36:103–113, 2017.

[73] X. Gao, S. Lin, and T. Y. Wong. Automatic Feature Learning to Grade Nuclear Cataracts Based on Deep Learning. *IEEE Transactions on Biomedical Engineering*, 62(11):2693–2701, 2015.

[74] X. W. Gao, R. Hui, and Z. Tian. Classification of CT Brain Images Based on Deep Learning Networks. *Computer Methods and Programs in Biomedicine*, 138:49–56, 2017.

[75] P. J. García-Laencina, J. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern Classification with Missing Data: A Review. *Neural Computing & Applications*, 19(2):263–282, 2010.

[76] L. A. Gatys, A. S. Ecker, and M. Bethge. Image Style Transfer using Convolutional Neural Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

[77] M. Gavaghan. Cardiac Anatomy and Physiology: A Review. *AORN Journal*, 67(4):800–822, 1998.

[78] D. T. Ginat, M. W. Fong, D. J. Tuttle, S. K. Hobbs, and R. C. Vyas. Cardiac Imaging: Part 1, MR Pulse Sequences, Imaging Planes, and Basic Anatomy. *American Journal of Roentgenology*, 197(4):808–815, 2011.

[79] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[80] I. Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv preprint arXiv:1701.00160*, 2016.

[81] I. Goodfellow, J. Pouget, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

[82] H. Greenspan, B. Van Ginneken, and R. M. Summers. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of An Exciting New Technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

[83] M. Haddadpour, S. Daneshvar, and H. Seyedarabi. PET and MRI Image Fusion Based on Combination of 2D Hilbert Transform and IHS Method. *Biomedical Journal*, 40(4):219–225, 2017.

[84] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P. M. Jodoin, and H. Larochelle. Brain Tumor Segmentation with Deep Neural Networks. *Medical Image Analysis*, 35:18–31, 2017.

[85] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[86] K. M He, X. Y Zhang, S. Q. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[87] L. He, D. Tao, X. Li, and X. Gao. Sparse Representation for Blind Image Quality Assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1146–1153, 2012.

[88] C. B. Higgins and A. de Roos. *MRI and CT of the Cardiovascular System*. Lippincott Williams & Wilkins, 2006.

[89] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[90] R. Hoffmann, F. Bertelshofer, C. Siegl, R. Janka, R. Grosso, and G. Greiner. Automated Heart Localization in Cardiac Cine MR Data. In *Bildverarbeitung für die Medizin*, pages 116–121, 2016.

[91] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285, 2016.

[92] A. Hoogi, A. Subramaniam, R. Veerapaneni, and D. L. Rubin. Adaptive Estimation of Active Contour Parameters using Convolutional Neural Networks and Texture Analysis. *IEEE Transactions on Medical Imaging*, 36(3):781–791, 2017.

[93] E. Hosseini-Asl, G. Gimel'farb, and A. El-Baz. Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3D Convolutional Network. *arXiv preprint arXiv:1607.00556*, 2016.

[94] J. Hu, L. Shen, and G. Sun. Squeeze-and-Excitation Networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[95] F.J. Huang, Y.-L. Boureau, and Y. LeCun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[96] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 3, 2017.

[97] Y. W. Huang, L. Shao, and A. F. Frangi. Simultaneous Super-Resolution and Cross-Modality Synthesis of 3D Medical Images using Weakly-Supervised Joint Convolutional Sparse Coding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6070–6079, 2017.

[98] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-Level Accuracy with 50x Fewer Parameters and 0.5 MB Model Size. *arXiv preprint arXiv:1602.07360*, 2016.

[99] S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*, 2015.

[100] C. Ionescu, O. Vantzos, and C. Sminchisescu. Matrix Backpropagation for Deep Networks with Structured Layers. In *IEEE International Conference on Computer Vision*, pages 2965–2973, 2015.

[101] O. Ishaq, S. K. Sadanandan, and C. Wählby. Deep Fish: Deep Learning Based Classification of Zebrafish Deformation for High-Throughput Screening. *SLAS DISCOVERY: Advancing Life Sciences R&D*, 22(1):102–107, 2017.

[102] P. Isola, J. Y. Zhu, T. H. Zhou, and A. A. Efros. Image-to-image Translation with Conditional Adversarial Networks. *arXiv preprint*, 2017.

[103] S. Ji, W. Xu, M. Yang, and K. Yu. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[104] Z. Jiao, X. Gao, Y. Wang, and J. Li. A Deep Feature Based Framework for Breast Masses Classification. *Neurocomputing*, 197:221–231, 2016.

[105] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[106] K. J. Jung. Synthesis Methods of Multiple Phase-cycled SSFP Images to Reduce the Band Artifact and Noise More Reliably. *Magnetic Resonance Imaging*, 28(1):103–118, 2010.

[107] M. Kallenberg, K. Petersen, M. Nielsen, A. Y. Ng, P. Diao, C. Igel, C. M. Vachon, K. Holland, R. R. Winkel, N. Karssemeijer, et al. Unsupervised Deep Learning Applied to Breast Density Segmentation and Mammographic Risk Scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, 2016.

[108] K. Kamnitsas, L. Chen, C. Ledig, D. Rueckert, and B. Glocker. Multi-scale 3D Convolutional Neural Networks for Lesion Segmentation in Brain MRI. In *Ischemic Stroke Lesion Segmentation*, volume 13, 2015.

[109] K. Kamnitsas, C. Ledig, V. FJ. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient Multi-scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation. *Medical Image Analysis*, 36:61–78, 2017.

[110] L. Kang, P. Ye, Y. Li, and D. Doermann. Convolutional Neural Networks for No-reference Image Quality Assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733–1740, 2014.

[111] K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun. Learning Convolutional Feature Hierarchies for Visual Recognition. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2010.

[112] J. Kawahara and G. Hamarneh. Multi-resolution-tract CNN with Hybrid Pretrained and Skin-lesion Trained Layers. In *International Workshop on Machine Learning in Medical Imaging*, pages 164–171. Springer, 2016.

[113] D. S. Kermany, M. Goldbaum, W. J Cai, C. C. Valentim, H. Y Liang, S. L. Baxter, A. McKeown, G. Yang, X. K Wu, F. B. Yan, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-based Deep Learning. *Cell*, 172(5):1122–1131, 2018.

[114] V. Klinke, S. Muzzarelli, N. Lauriers, D. Locca, G. Vincenti, P. Monney, C. Lu, D. Nothnagel, G. Pilz, M. Lombardi, et al. Quality Assessment of Cardiovascular Magnetic Resonance in the Setting of the European CMR Registry: Description and Validation of Standardized Criteria. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1, 2013.

[115] B. Kong, Y. Zhan, M. Shin, T. Denny, and S. Zhang. Recognizing End-diastole and End-systole Frames via Deep Temporal Regression Network. In *Medical Image Computing and Computer-assisted Intervention*, pages 264–272. Springer, 2016.

[116] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, and N. Karssemeijer. Large Scale Deep Learning for Computer Aided Detection of Mammographic Lesions. *Medical Image Analysis*, 35:303–312, 2017.

[117] C. M. Kramer, J. Barkhausen, S. D. Flamm, R. J. Kim, and E. Nagel. Standardized Cardiovascular Magnetic Resonance (CMR) Protocols 2013 Update. *Journal of Cardiovascular Magnetic Resonance*, 15(1):91, 2013.

[118] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems*, pages 1097–1105, 2012.

[119] K. Krupa and M. Bekiesińska-Figatowska. Artifacts in Magnetic Resonance Imaging. *Polish Journal of Radiology*, 80:93, 2015.

[120] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable Visual Attributes for Face Verification and Image Search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[121] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi. A Dataset and A Technique for Generalized Nuclear Segmentation for Computational Pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017.

[122] J. Lambert, O. Sener, and S. Savarese. Deep Learning under Privileged Information Using Heteroscedastic Dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2018.

[123] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436, 2015.

[124] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[125] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 609–616. ACM, 2009.

[126] S. E. Lee, C. Nguyen, Y. B. Xie, Z. X Deng, Z. W Zhou, D. B Li, and H. J Chang. Recent Advances in Cardiac Magnetic Resonance Imaging. *Korean Circulation Journal*, 49(2):146–159, 2019.

[127] J. Lei, G. Li, D. Tu, and Q. Guo. Convolutional Restricted Boltzmann Machines Learning for Robust Visual Tracking. *Neural Computing and Applications*, 25(6):1383–1391, 2014.

[128] A. Li, J. Cheng, D. W. K. Wong, and J. Liu. Integrating Holistic and Local Deep Features for Glaucoma Classification. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 1328–1331. IEEE, 2016.

[129] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, and T. Wang. A Cross-Modality Learning Approach for Vessel Segmentation in Retinal Images. *IEEE Transactions on Medical Imaging*, 35(1):109–118, 2016.

[130] S. T Li, X. D. Kang, L. Y. Fang, J. W. Hu, and H. T Yin. Pixel-level Image Fusion: A Survey of the State of the Art. *Information Fusion*, 33:100–112, 2017.

[131] Y. Li, W. Liang, Y. Zhang, H. An, and J. Tan. Automatic Lumbar Vertebrae Detection Based on Feature Fusion Deep Learning for Partial Occluded C-arm X-ray Images. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 647–650. IEEE, 2016.

[132] J. Lieman-Sifry, M. Le, F. Lau, S. Sall, and D. Golden. FastVentricle: Cardiac Segmentation with ENet. In *International Conference on Functional Imaging and Modeling of the Heart*, pages 127–138. Springer, 2017.

[133] L. Lin, G. R. Wang, W. M. Zuo, X. C. Feng, and L. Zhang. Cross-domain Visual Matching via Generalized Similarity Measure and Feature Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1089–1102, 2017.

[134] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo. Person re-identification by iterative re-weighted sparse ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1629–1642, 2015.

[135] P. Liskowski and K. Krawiec. Segmenting Retinal Blood Vessels with Deep Neural Networks. *IEEE Transactions on Medical Imaging*, 35(11):2369–2380, 2016.

[136] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. Van Ginneken, and C. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017.

[137] F. Liu, C. Shen, and G. Lin. Deep Convolutional Neural Fields for Depth Estimation From A Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[138] G. Liu, H. Bao, and B. Han. A Stacked Autoencoder-Based Deep Neural Network for Achieving Gearbox Fault Diagnosis. *Mathematical Problems in Engineering*, 2018, 2018.

[139] H. Liu and Z. Wang. Perceptual Quality Assessment of Medical Images. In *Encyclopedia of Biomedical Engineering*. Elsevier, 2019.

[140] M. Y. Liu and O. Tuzel. Coupled Generative Adversarial Networks. In *Conference on Neural Information Processing Systems*, pages 469–477, 2016.

[141] S. Liu, S. Liu, W. Cai, H. Che, S. Pujol, R. Kikinis, D. Feng, M. J. Fulham, et al. Multimodal Neuroimaging Feature Learning for Multiclass Diagnosis of Alzheimer's Disease. *IEEE Transactions on Biomedical Engineering*, 62(4):1132–1140, 2015.

[142] S. Lo, S. Lou, J. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun. Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection. *IEEE Transactions on Medical Imaging*, 14(4):711–718, 1995.

[143] E. Long, H. T. Lin, Z. Z Liu, X. H. Wu, L. M Wang, J. W Jiang, Y. Y An, Z. L Lin, X. Y. Li, J. J. Chen, et al. An Artificial Intelligence Platform for the Multihospital Collaborative Management of Congenital Cataracts. *Nature Biomedical Engineering*, 1(2):0024, 2017.

[144] J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[145] X. Lu, D. Xu, and D. Liu. Robust 3D Organ Localization with Dual Learning Architectures and Fusion. In *International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, pages 12–20. Springer, 2016.

[146] X. G. Lu and M. P. Jolly. Discriminative Context Modeling using Auxiliary Markers for LV Landmark Detection From A Single MR Image. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 105–114. Springer, 2012.

[147] X. G Lu, M. P. Jolly, B. Georgescu, C. Hayes, P. Speier, M. Schmidt, X. M. Bi, R. Kroeker, D. Comaniciu, P. Kellman, et al. Automatic view planning for cardiac mri acquisition. In *Medical Image Computing and Computer-Assisted Intervention*, pages 479–486. Springer, 2011.

[148] Y. Lu, K. A. Connelly, A. J. Dick, G. A. Wright, and P. E Radau. Watershed Segmentation of Basal Left Ventricle for Quantitation of Cine Cardiac MRI Function. *Journal of Cardiovascular Magnetic Resonance*, 13(S1):P4, 2011.

[149] Y. Lu, P. Radau, K. Connelly, A. Dick, and G. Wright. Pattern Recognition of Abnormal Left Ventricle Wall Motion in Cardiac MR. In *International Conferenceon Medical Image Computing and Computer-Assisted Intervention*, pages 750–758, Springer, 2009.

[150] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[151] L. Y. Ma, X. K. Yang, and D. C. Tao. Person Re-identification Over Camera Networks Using Multi-task Distance Metric Learning. *IEEE Transactions on Image Processing*, 23(8):3656–3670, 2014.

[152] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In *International Conference on Machine Learning*, pages 1–8, 2013.

[153] An. Madabhushi and G. Lee. Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities, 2016.

[154] D. Mahapatra. Landmark Detection in Cardiac MRI Using Learned Local Image Statistics. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 115–124. Springer, 2012.

[155] B. Mahasseni and S. Todorovic. Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3054–3062, 2016.

[156] J.T. Marcus, M.J. Götte, L.K. DeWaal, M.R. Stam, R.J. Van der Geest, R.M. Heethaar, and A.C. Van Rossum. The Influence of Through-Plane Motion on Left

Ventricular Volumes Measured by Magnetic Resonance Imaging: Implications for Image Acquisition and Analysis. *Journal of Cardiovascular Magnetic Resonance*, 1:1–6, 1999.

[157] A. Masood, A. Al-Jumaily, and K. Anam. Self-Supervised Learning Model for Skin Cancer Diagnosis. In *International IEEE/EMBS Conference on Neural Engineering*, pages 1012–1015. IEEE, 2015.

[158] J. J. McMurray, S. Adamopoulos, S. D. Anker, A. Auricchio, M. Böhm, K. Dickstein, V. Falk, G. Filippatos, C. Fonseca, et al. ESC Guidelines for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association (HFA) of the ESC. *European Journal of Heart Failure*, 14(8):803–869, 2012.

[159] S. Miao, Z Jane Wang, and R. Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE Transactions on Medical Imaging*, 35(5):1352–1363, 2016.

[160] MICCAI. 2017 ACDC Challenge. `https://www.creatis.insa-lyon.fr/Challenge/acdc/`. Accessed 25 Oct 2017.

[161] C.A. Miller, P. Jordan, A. Borg, R. Argyle, D. Clark, K. Pearce, and M. Schmitt. Quantification of Left Ventricular Indices from SSFP Cine Imaging: Impact of Real-world Variability in Analysis Methodology and Utility of Geometric Modeling. *Journal of Cardiovascular Magnetic Resonance*, 37:1213–1222, 2013.

[162] M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[163] A. A. Mohamed, W. A. Berg, H. Peng, Y. H. Luo, R. C. Jankowitz, and S. D. Wu. A Deep Learning Method for Classifying Mammographic Breast Density Categories. *Medical Physics*, 45(1):314–321, 2018.

[164] A. K. Moorthy and A. C. Bovik. Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality. *IEEE Transactions on Image Processing*, 20:3350–3364, 2011.

[165] A. Mortazi, J. Burt, and U. Bagci. Multi-Planar Deep Segmentation Networks for Cardiac Substructures from MRI and CT. *arXiv preprint arXiv:1708.00983*, 2017.

[166] I. Myrtveit, E. Stensrud, and U. H. Olsson. Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-based Methods. *IEEE Transactions on Software Engineering*, 27(11):999–1013, 2001.

[167] M. P. Nash and P. J. Hunter. Computational Mechanics of the Heart. *Journal of Elasticity and the Physical Science of Solids*, 61(1-3):113–141, 2000.

[168] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. R. Soroushmehr, K. Ward, M. H. Jafari, B. Felfeliyan, B. Nallamothu, and K. Najarian. Vessel Extraction in X-ray Angiograms using Deep Learning. In *2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society*, pages 643–646. IEEE, 2016.

[169] T. A. Ngo, Z. Lu, and G. Carneiro. Combining Deep Learning and Level Set for the Automated Segmentation of the Left Ventricle of the Heart From Cardiac Cine Magnetic Resonance. *Medical Image Analysis*, 35:159–171, 2017.

[170] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. G Shen. Medical Image Synthesis with Context-aware Generative Adversarial Networks. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 417–425. Springer, 2017.

[171] D. Nie, H. Zhang, E. Adeli, L. Liu, and D. Shen. 3D Deep Learning for Multimodal Imaging-guided Survival Time Prediction of Brain Tumor Patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–220. Springer, 2016.

[172] Z. X. Niu, M. Zhou, L. Wang, X. B. Gao, and G. Hua. Ordinal Regression with Multiple Output CNN for Age Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016.

[173] A. Odena. Semi-supervised Learning with Generative Adversarial Networks. *arXiv preprint arXiv:1606.01583*, 2016.

[174] A. Odena, C. Olah, and J. Shlens. Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 2642–2651. JMLR. org, 2017.

[175] L. H. Opie. *Heart Physiology: From Cell to Circulation*. Lippincott Williams & Wilkins, 2004.

[176] World Health Organisation. Cardiovascular Diseases (CVDs) Fact Sheet. `http://www.who.int/mediacentre/factsheets/fs317/en/`. Accessed 11 July 2017.

[177] M. A. Oskoei and H. Hu. A Survey on Edge Detection Methods. *University of Essex, UK*, 2010.

[178] M. Paknezhad, S. Marchesseau, and M. S. Brown. Automatic Basal Slice Detection for Cardiac Analysis. *Journal of Medical Imaging*, 3(3):034004–034004, 2016.

[179] A. Payan and G. Montana. Predicting Alzheimer's Disease: A Neuroimaging Study with 3D Convolutional Neural Networks. *arXiv preprint arXiv:1502.02506*, 2015.

[180] O. AB. Penatti, K. Nogueira, and J. A. dos Santos. Do Deep Features Generalize From Everyday Objects to Remote Sensing and Aerial Scenes Domains? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition workshops*, pages 44–51, 2015.

[181] S. Pereira, A. Pinto, V. Alves, and C. A. Silva. Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. *IEEE Transactions on Medical Imaging*, 35(5):1240–1251, 2016.

[182] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, et al. Imaging in Population Science: Cardiovascular Magnetic Resonance in 100,000 Participants of UK Biobank-Rationale, Challenges and Approaches. *Journal of Cardiovascular Magnetic Resonance*, 15(1):46, 2013.

[183] S. E. Petersen, P. M. Matthews, J. M. Francis, M. D. Robson, F. Zemrak, R. Boubertakh, A. A. Young, S. Hudson, P. Weale, S. Garratt, et al. UK Biobanks Cardiovascular Magnetic Resonance Protocol. *Journal of Cardiovascular Magnetic Resonance*, 18(1):8, 2015.

[184] S.E. Petersen, N. Aung, M.M. Sanghvi, F. Zemrak, K. Fung, J.M. Paiva, J.M. Francis, M.Y. Khanji, E. Lukaschuk, A.M. Lee, V. Carapella, Y.J. Kim, P. Leeson, S.K. Piechnik, and S. Neubauer. Reference Ranges for Cardiac Structure and Function Using Cardiovascular Magnetic Resonance (CMR) in Caucasians from the UK Biobank Population Cohort. *Journal of Cardiovascular Magnetic Resonance*, 19:18, 2017.

[185] Ben Poole, Alexander A Alemi, Jascha Sohl-Dickstein, and Anelia Angelova. Improved generator objectives for GANs. *arXiv preprint arXiv:1612.02780*, 2016.

[186] K. Pradeep, S. Balasubramanian, H. Karnan, and K. K. Babu. Segmentation of Fused CT and MRI Images with Brain Tumor. *Asian Journal of Science and Applied Technology*, 6(1):1–4, 2017.

[187] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen. Deep Feature Learning for Knee Cartilage Segmentation Using A Triplanar Convolutional Neural Network. In *International Conferenceon Medical Image Computing and Computer-Assisted Intervention*, pages 246–253, Springer, 2013.

[188] E. Pusey, R. B. Lufkin, R. Brown, M. A. Solomon, D. D. Stark, R. Tarr, and W. Hanafee. Magnetic Resonance Imaging Artifacts: Mechanism and Clinical Significance. *Radiographics*, 6(5):891–911, 1986.

[189] T. Qaiser, K. Sirinukunwattana, K. Nakane, Y. W. Tsang, D. Epstein, and N. Rajpoot. Persistent Homology for Fast Tumor Segmentation in Whole Slide Histology Images. *Procedia Computer Science*, 90:119–124, 2016.

[190] G. Quellec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard. Deep Image Mining for Diabetic Retinopathy Screening. *Medical Image Analysis*, 39:178–193, 2017.

[191] S. Reed, Z. Akata, X. C Yan, L. Logeswaran, B. Schiele, and H. L Lee. Generative Adversarial Text to Image Synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[192] C. Rickers, N. M. Wilke, M. Jerosch-Herold, S. A. Casey, P. Panse, N. Panse, J. Weil, A. G. Zenovich, and B. J. Maron. Utility of Cardiac Magnetic Resonance Imaging in the Diagnosis of Hypertrophic Cardiomyopathy. *Circulation*, 112(6):855–861, 2005.

[193] J. P. Ridgway. Cardiovascular Magnetic Resonance Physics for Clinicians: Part I. *Journal of Cardiovascular Magnetic Resonance*, 12(1):1, 2010.

[194] D. Ripley, T. Musa, L. Dobson, S. Plein, and J. Greenwood. Cardiovascular Magnetic Resonance Imaging: What the General Cardiologist Should Know. *Heart*, 102(19):1589–1603, 2016.

[195] H. R. Roth, A. Farag, L. Lu, E. B. Turkbey, and R. M. Summers. Deep Convolutional Networks for Pancreas Segmentation in CT Imaging. In *Medical Imaging 2015: Image Processing*, volume 9413, page 94131G. International Society for Optics and Photonics, 2015.

[196] H. R. Roth, L. Lu, J. Liu, J. Yao, A. Seff, K. Cherry, L. Kim, and R. M. Summers. Improving Computer-aided Detection using Convolutional Neural Networks and Random View Aggregation. *IEEE Transactions on Medical Imaging*, 35(5):1170–1181, 2016.

[197] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers. A New 2.5D Representation for Lymph Node Detection using Random Sets of Deep Convolutional Neural Network Observations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–527. Springer, 2014.

[198] H. R. Roth, J. Yao, L. Lu, J. Stieger, J. E. Burns, and R. M. Summers. Detection of Sclerotic Spine Metastases via Random Aggregation of Deep Convolutional Neural Network Classifications. In *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, pages 3–12. Springer, 2015.

[199] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Transactions on Graphics*, 23(3):309–314, 2004.

[200] D. Rueckert, B. Glocker, and B. Kainz. Learning Clinically Useful Information From Images: Past, Present and Future, 2016.

[201] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Representations by Back-propagating Errors. *Cognitive Modeling*, 5(3):1, 1988.

[202] M. A. Saad, A. C. Bovik, and C. Charrier. Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.

[203] A.A. Salah, E. Alpaydin, and L. Akarun. A Selective Attention-based Method for Visual Pattern Recognition with Application to Handwritten Digit Recognition and Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:420–425, 2002.

[204] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *Conference on Neural Information Processing Systems*, pages 2234–2242, 2016.

[205] D. Sarikaya, J. J. Corso, and K. A. Guru. Detection and Localization of Robotic Tools in Robot-Assisted Surgery Videos Using Deep Neural Networks for Region Proposal and Detection. *IEEE Transactions on Medical Imaging*, 36(7):1542–1549, 2017.

[206] G. Schillaci, P. Verdecchia, C. Porcellati, O. Cuccurullo, C. Cosco, and F. Perticone. Continuous Relation Between Left Ventricular Mass and Cardiovascular Risk in Essential Hypertension. *Hypertension*, 35(2):580–586, 2000.

[207] T. Schlegl, J. Ofner, and G. Langs. Unsupervised Pre-training Across Image Domains Improves Lung Tissue Classification. In *International MICCAI Workshop on Medical Computer Vision*, pages 82–93. Springer, 2014.

[208] G. L. Schlomer, S. Bauman, and N. A. Card. Best Practices for Missing Data Management in Counseling Psychology. *Journal of Counseling Psychology*, 57(1):1, 2010.

[209] J. Schulz-Menger, D. A. Bluemke, J. Bremerich, S. D. Flamm, M. A. Fogel, M. G. Friedrich, R. J. Kim, F. von Knobelsdorff-Brenkenhoff, C. M. Kramer, D. J. Pennell, et al. Standardized Image Interpretation and Post Processing in Cardiovascular Magnetic Resonance: Society for Cardiovascular Magnetic Resonance (SCMR) Board of Trustees Task Force on Standardized Post Processing. *Journal of Cardiovascular Magnetic Resonance*, 15(1):35, 2013.

[210] P. Sermanet, S. Chintala, and Y. LeCun. Convolutional Neural Networks Applied to House Numbers Digit Classification. In *International Conference on Pattern Recognition*, pages 3288–3291, IEEE, 2012.

[211] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. Van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[212] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition Workshops, pages 806–813, 2014.

[213] V. Sharmanska and N. Quadrianto. Learning From the Mistakes of Others: Matching Errors in Cross-Dataset Learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3967–3975, 2016.

[214] D. Shen, G. Wu, and H. I. Suk. Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, 19:221–248, 2017.

[215] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang. Stacked Deep Polynomial Network Based Representation Learning for Tumor Classification with Small Ultrasound Image Dataset. *Neurocomputing*, 194:87–94, 2016.

[216] H. Shin, M. R. Orton, D. J. Collins, S. J. Doran, and M. O. Leach. Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in A Pilot Study Using 4D Patient Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2012.

[217] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587):484, 2016.

[218] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-scale Image Recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[219] K. Sirinukunwattana, S. E. Ahmed Raza, Y. W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot. Locality Sensitive Deep Learning for Detection and Classification of Nuclei in Routine Colon Cancer Histology Images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016.

[220] Y. Song, E. Tan, X. Jiang, J. Z. Cheng, D. Ni, S. Chen, B. Lei, and T. Wang. Accurate Cervical Cell Segmentation From Overlapping Clumps in Pap Smear Images. *IEEE Transactions on Medical Imaging*, 36(1):288–300, 2017.

[221] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang. Accurate Segmentation of Cervical Cytoplasm and Nuclei Based on Multiscale Convolutional Network and Graph Partitioning. *IEEE Transactions on Biomedical Engineering*, 62(10):2421–2433, 2015.

[222] C Spampinato, S Palazzo, D Giordano, M Aldinucci, and R Leonardi. Deep Learning for Automated Skeletal Bone Age Assessment in X-ray Images. *Medical Image Analysis*, 36:41–51, 2017.

[223] S. Standring. *Gray's Anatomy E-book: the Anatomical Basis of Clinical Practice*. Elsevier Health Sciences, 2015.

[224] M. F. Stollenga, W. Byeon, M. Liwicki, and J. Schmidhuber. Parallel Multi-dimensional LSTM, with Application to Fast Biomedical Volumetric Image Segmentation. In *Advances in Neural Information Processing Systems*, pages 2998–3006, 2015.

[225] H. Suk, S. Lee, D. Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Hierarchical Feature Representation and Multimodal Fusion with Deep Learning for AD/MCI Diagnosis. *NeuroImage*, 101:569–582, 2014.

[226] H. Suk, S. Lee, D. Shen, Alzheimers Disease Neuroimaging Initiative, et al. Latent Feature Representation with Stacked Auto-encoder for AD/MCI Diagnosis. *Brain Structure and Function*, 220(2):841–859, 2015.

[227] M. J. Sutton and N. Sharpe. Left Ventricular Remodeling After Myocardial Infarction: Pathophysiology and Therapy. *Circulation*, 101(25):2981–2988, 2000.

[228] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In *Association for the Advancement of Artificial Intelligence*, volume 4, page 12, 2017.

[229] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[230] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing Properties of Neural Networks. *arXiv preprint arXiv:1312.6199*, 2013.

[231] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[232] N. Tajbakhsh and K. Suzuki. Comparing Two Classes of End-to-End Machine Learning Models in Lung Nodule Detection and Classification: MTANNs vs. CNNs. *Pattern Recognition*, 63:476–486, 2017.

[233] L. K. Tan, Y. M. Liew, E. Lim, and R. A. McLaughlin. Convolutional Neural Network Regression for Short-Axis Left Ventricle Segmentation in Cardiac Cine MR Sequences. *Medical Image Analysis*, 39:78–86, 2017.

[234] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki. Automated Detection of Pulmonary Nodules in PET/CT Images: Ensemble False-Positive Reduction Using a Convolutional Neural Network Technique. *Medical Physics*, 43(6Part1):2821–2827, 2016.

[235] A. Thompson and N. Maredia. Cardiovascular Magnetic Resonance Imaging for the Assessment of Ischemic Heart Disease. *Continuing Cardiology Education*, 3(2):56–63, 2017.

[236] K. H. Thung, C. Y. Wee, P. T. Yap, D. Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Neurodegenerative Disease Diagnosis Using Incomplete Multi-Modality Data via Matrix Shrinkage and Completion. *NeuroImage*, 91:386–400, 2014.

[237] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[238] P. V. Tran. A Fully Convolutional Neural Network for Cardiac Segmentation in Short-Axis MRI. *arXiv preprint arXiv:1604.00494*, 2016.

[239] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy. Endonet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.

[240] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous Deep Transfer Across Domains and Tasks. In *IEEE Conference on Computer Vision*, pages 4068–4076. IEEE, 2015.

[241] G. Urban, M. Bendszus, F. Hamprecht, and J. Kleesiek. Multi-Modal Brain Tumor Segmentation Using Deep Convolutional Neural Networks. *MICCAI BraTS (Brain Tumor Segmentation) Challenge. Proceedings, winning Contribution*, pages 31–35, 2014.

[242] A. van der Graaf, P. Bhagirath, S. Ghoerbien, and M. Götte. Cardiac Magnetic Resonance Imaging: Artefacts for Clinicians. *Netherlands Heart Journal*, 22(12):542–549, 2014.

[243] B. Van Ginneken, A. AA. Setio, C. Jacobs, and F. Ciompi. Off-the-Shelf Convolutional Neural Network Features for Pulmonary Nodule Detection in Computed Tomography Scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging*, pages 286–289. IEEE, 2015.

[244] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez. Fast Convolutional Neural Network Training Using Selective Data Sampling: Application to Hemorrhage Detection in Color Fundus Images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, 2016.

[245] G. van Tulder and M. de Bruijne. Combining Generative and Discriminative Representation Learning for Lung CT Analysis with Convolutional Restricted Boltzmann Machines. *IEEE Transactions on Medical Imaging*, 35(5):1262–1272, 2016.

[246] V. Vapnik and A. Vashist. A New Learning Paradigm: Learning Using Privileged Information. *Neural Network*, 22(5-6):544–557, 2009.

[247] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in A Deep Network with A Local Denoising Criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[248] C. Wang, X. Yan, M. Smith, K. Kochhar, M. Rubin, S. M. Warren, J. Wrobel, and H. Lee. A Unified Framework for Automatic Wound Segmentation and Analysis with Deep Convolutional Neural Networks. In *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pages 2415–2418. IEEE, 2015.

[249] J. Wang, H. Ding, F. A. Bidgoli, B. Zhou, C. Iribarren, S. Molloi, and P. Baldi. Detecting Cardiovascular Disease from Mammograms With Deep Learning. *IEEE Transactions on Medical Imaging*, 36(5):1172–1181, 2017.

[250] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu. CNN-RNN: An Unified Framework for Multi-label Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016.

[251] S. Wang, Y. Yin, G. Cao, B. Wei, Y. Zheng, and G. Yang. Hierarchical Retinal Blood Vessel Segmentation Based on Feature and Ensemble Learning. *Neurocomputing*, 149:708–717, 2015.

[252] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end Text Recognition with Convolutional Neural Networks. In *2012 21st International Conference on Pattern Recognition*, pages 3304–3308. IEEE, 2012.

[253] Z. Wang, G. X. Wu, H. R. Sheikh, E. P. Simoncelli, E. H. Yang, and A. C. Bovik. Quality Aware Images. *IEEE Transactions on Image Processing*, 15(6):1680–1689, 2006.

[254] J. Weese and C. Lorenz. Four Challenges in Medical Image Analysis From An Industrial Perspective, 2016.

[255] W. M. Wells III. Medical Image Analysis–Past, Present, and Future, 2016.

[256] D. Williams, X. J Liao, Y. Xue, L. Carin, and B. Krishnapuram. On Classification with Incomplete Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:427–436, 2007.

[257] A. Wu, Z. Xu, M. Gao, M. Buty, and D. J. Mollura. Deep Vessel Tracking: A Generalized Probabilistic Approach via Deep Learning. In *IEEE International Symposium on Biomedical Imaging*, pages 1363–1367. IEEE, 2016.

[258] C. P. Wu, W. Wen, T. Afzal, Y. M. Zhang, Y. R. Chen, and H. Li. A Compact DNN: Approaching GoogleNet-Level Accuracy of Classification and Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[259] J. Xie, G. Dai, F. Zhu, E. Wong, and Y. Fang. DeepShape: Deep-Learned Shape Descriptor for 3D Shape Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[260] F. Xing, Y. Xie, and L. Yang. An Automatic Learning-Based Framework for Robust Nucleus Segmentation. *IEEE Transactions on Medical Imaging*, 35(2):550–566, 2016.

[261] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi. Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images. *IEEE Transactions on Medical Imaging*, 35(1):119–130, 2016.

[262] Y. Xu, Z. Jia, Y. Ai, F. Zhang, M. Lai, I. Eric, and C. Chang. Deep Convolutional Activation Features for Large Scale Brain Tumor Histopathology Image Classification and Segmentation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 947–951. IEEE, 2015.

[263] Y. Xu, T. Mo, Q. Feng, P. Zhong, M. Lai, I. Eric, and C. Chang. Deep Learning of Feature Representation with Multiple Instance Learning for Medical Image Analysis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1626–1630. IEEE, 2014.

[264] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng. Blind Image Quality Assessment Using Joint Statistics of Gradient Magnitude and Laplacian Features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014.

[265] D. Yang, S. Zhang, Z. Yan, C. Tan, K. Li, and D. Metaxas. Automated Anatomical Landmark Detection Ondistal Femur Surface Using Convolutional Neural Network. In *2015 IEEE 12th International Symposium on Biomedical Imaging*, pages 17–21. IEEE, 2015.

[266] M. Yang, L. Zhang, X.D. Feng, and D. Zhang. Sparse Representation Based Fisher Discrimination Dictionary Learning for Image Classification. *International Journal of Computer Vision*, 109:209–232, 2014.

[267] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A Heng. Automated Melanoma Recognition in Dermoscopy Images via Very Deep Residual Networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017.

[268] L. T. Yu, W. N. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Association for the Advancement of Artificial Intelligence*, 2017.

[269] X. T. Yuan, X. B. Liu, and S. C. Yan. Visual Classification with Multitask Joint Sparse Representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012.

[270] J. Zhang. Big Bata Issues in Medical Imaging Informatics. In *Medical Imaging 2015: PACS and Imaging Informatics: Next Generation and Innovations*, volume 9418, page 941803. International Society for Optics and Photonics, 2015.

[271] L. Zhang, A. Gooya, B. Dong, R. Hua, S. E. Petersen, P. Medrano-Gracia, and A. F. Frangi. Automated Quality Assessment of Cardiac MR Images Using Convolutional Neural Networks. In *International Workshop on Simulation and Synthesis in Medical Imaging*, pages 138–145. Springer, 2016.

[272] L. Zhang, A. Gooya, and A. F. Frangi. Semi-supervised Assessment of Incomplete LV Coverage in Cardiac MRI Using Generative Adversarial Nets. In *MICCAI Workshop on Simulation and Synthesis in Medical Imaging*, pages 61–68. Springer, 2017.

[273] L. Zhang, A. Gooya, M. Pereanez, B. Dong, S. Piechnik, S. Neubauer, S. Petersen, and A. F. Frangi. Automatic Assessment of Full Left Ventricular Coverage in Cardiac Cine Magnetic Resonance Imaging with Fisher Discriminative 3D CNN. *IEEE Transactions on Biomedical Engineering*, 2018.

[274] L. Zhang, M. Pereañez, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi. Multi-input and Dataset-Invariant Adversarial Learning (MDAL) for Left and Right-Ventricular Coverage Estimation in Cardiac MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 481–489. Springer, 2018.

[275] Q. Zhang, Y. Xiao, W. Dai, J. Suo, C. Wang, J. Shi, and H. Zheng. Deep Learning Based Classification of Breast Tumors with Shear-wave Elastography. *Ultrasonics*, 72:150–157, 2016.

[276] S. Zhang and D. Metaxas. Large-scale Medical Image Analytics: Recent Methodologies, Applications and Future Directions, 2016.

[277] T. Z. Zhang, B. Ghanem, S. Liu, and N. Ahuja. Robust Visual Tracking via Structured Multi-task Sparse Learning. *International Journal of Computer Vision*, 101(2):367–383, 2013.

[278] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen. Deep Convolutional Neural Networks for Multi-modality Isointense Infant Brain Image Segmentation. *NeuroImage*, 108:214–224, 2015.

[279] X. T. Zhen, Z. J. Wang, A. Islam, M. Bhaduri, I. Chan, and S. Li. Multi-Scale Deep Networks and Regression Forests for Direct Bi-ventricular Volume Estimation. *Medical Image Analysis*, 30:120–129, 2016.

[280] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *arXiv preprint*, 2017.

[281] D. Zikic, Y. Ioannou, M. Brown, and A. Criminisi. Segmentation of Brain Tumor Tissues with Convolutional Neural Networks. *MICCAI Multimodal Brain Tumor Segmentation Challenge*, pages 36–39, 2014.