

Imperial College of Science, Technology and Medicine
Department of Computing

Learning Anatomical Image Representations for Cardiac Imaging

Ozan Oktay

A dissertation submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London

December 2017

Abstract

Advances in cardiac imaging techniques play a vital role in fighting against cardiovascular diseases by helping achieve early and accurate diagnoses as well as guidance in interventional cardiac procedures. Along with these imaging systems, computer-aided diagnosis (CAD) is essential in modern healthcare to yield useful and comprehensive clinical information in a short time. However, CAD systems have not been widely adopted in clinical practice due to technical limitations imposed by the imaging systems. In particular, cardiac cine Magnetic Resonance (MR) imaging has a low through-plane resolution and may contain motion artefacts due to the acquisition process. Similarly, the quality of cardiac ultrasound images is operator-dependent and often has poor signal-to-noise-ratio, which presents difficulties for automated image analysis.

To tackle these limitations and enhance the accuracy and robustness of the automated image analysis, this thesis focuses on the development and application of state-of-the-art machine learning (ML) techniques in cardiac multi-modal imaging. Specifically, we propose new image feature representation types that are learnt with ML models and aimed at highlighting the correspondences between multi-modal data. These representations are also intended to visualise the cardiac anatomy in more detail for better interpretation and analysis. Moreover, we investigate how quantitative analysis can benefit when these learnt image representations are used in segmentation, motion-tracking and multi-modal image registration.

Specifically, a probabilistic edge-map representation is introduced to identify anatomical correspondences in multi-modal cardiac images and demonstrate its use in spatial image alignment and anatomical landmark localisation problems. Additionally, a novel image super-resolution framework is introduced for the enhancement of cardiac cine MR images and we show that high resolution image representation can be useful and informative for various types of subsequent analysis including volumetric measurements and strain analysis.

Acknowledgements

First and foremost I wish to thank my supervisor Prof. Daniel Rueckert, 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 BioMedIA group over the last four years. I am very thankful to everyone, especially to Jose, Jonathan, Kevin, Gregor, Salim, Martin, and Giacomo for all the good and fun memories. Special thanks to Wenjia and Wenzhe for their help and guidance, without whom we would not have been able to complete the research work presented in this thesis.

I also need to thank to Dr. Declan O'Regan, Dr. Antonio de Marvao, and Dr. Timothy Dawes of the Robert Steiner MR imaging unit at Hammersmith Hospital for their clinical feedback and for providing almost all the MR imaging data presented in this thesis.

I also would like to express my special thanks to Agnes for her loving support and endless patience during my studies. Finally, I am deeply grateful to my parents and brother for their constant support and love, who have given me all the strength and courage to follow my dreams.

To my ever dearest mother ...

Declaration of originality

I declare that the work presented in this thesis is my own, unless specifically acknowledged.

Ozan Oktay.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Acronyms

ACNN Anatomically Constrained Neural Network.

AE Autoencoder.

AL Adversarial Learning.

AW Anterior Wall.

CAD Computer-Aided Diagnosis.

CMR Magnetic Resonance Imaging.

CNN Convolutional Neural Network.

CT Computed Tomography.

CVD Cardiovascular Disease.

DBN Deep Belief Network.

DF Decision Forest.

ED End-Diastolic.

EF Ejection Fraction.

EM Expectation Maximisation.

ES End-Systolic.

FFD Free-Form Deformations.

GM Gradient Magnitude.

HR High Resolution.

LAX Long Axis.

LR Low Resolution.

LV Left Ventricle.

MI Mutual Information.

ML Machine Learning.

MR Magnetic Resonance.

MRF Markov Random Field.

MRI Magnetic Resonance Imaging.

PEM Probabilistic Edge Map.

PEM-MA Probabilistic Edge Map based Multi-Atlas Segmentation.

PSNR Peak Signal-to-Noise Ratio.

SAX Short Axis.

SDF Structured Decision Forest.

SR Super Resolution.

SR-CNN Super-Resolution Convolutional Neural Network.

SSD Sum of Squared Differences.

SSFP Steady-State Free Precession.

SSIM Structural Similarity Index Measure.

SV Stroke Volume.

US Ultrasound.

Nomenclature

PEM Based Image Analysis

$\mathcal{D}$	Landmark distance offset
$\mathcal{F}$	Hough votemap space
$\mathcal{H}$	Shannon entropy
$\mathcal{I}$	Information gain
$\mathcal{M}$	Image meta-information
$\mathcal{P}$	Decision tree training samples
$\mathcal{X}$	Input image feature space
$\mathcal{Y}$	Decision tree output space
A	Global affine transformation
I	Target image
J	Source or atlas image
P	PEM representation of an image
T	FFD deformation parameters

Image Super-Resolution

θ	Neural network parameters
w	Convolution filters
$\mathcal{H}$	Auto-encoder latent space
$\mathcal{X}$	Low-resolution input space
$\mathcal{Y}$	High-resolution output space
Φ	Learnt mapping function
Ψ_{ℓ_1}	Huber loss function
f	Image encoder
g	Image decoder
h	CNN feature-maps
L	Training loss function
p	Latent space predictor

Contents

Abstract	i
Acronyms	vii
Nomenclature	x
1 Introduction	1
1.1 Clinical Background and Motivation	1
1.1.1 The Anatomy and Function of the Heart	1
1.1.2 The Cardiac Cycle and Electrical Activation	3
1.1.3 Cardiac Magnetic Resonance (CMR) Imaging	4
1.1.4 Indices of Cardiac Function	8
1.1.5 Cardiac Ultrasound Imaging	9
1.2 Technical Challenges in Multi-modal Cardiac Image Analysis and Objectives . .	11
1.3 Thesis Contributions	14
1.4 Thesis Overview	16
2 Review of Machine Learning Methods in Cardiac Image Analysis	18
2.1 Cardiac MR and Ultrasound Image Segmentation	19
2.1.1 Energy Minimisation Methods	20
2.1.2 Gaussian Mixture Models	21
2.1.3 Multi-Atlas Segmentation Methods	22
2.1.4 Anatomical Priors in Cardiac Segmentation	23

2.2	Super-Resolution in Magnetic Resonance Images	25
2.2.1	Variational Inverse Methods	27
2.2.2	Regression Models for Image Super-Resolution	28
2.3	Multi-Modal Cardiac Image Registration	29
2.3.1	Transformation Models and Optimisation Techniques	30
2.3.2	Image Similarity Criteria in Multi-Modal Image Registration	31
2.3.3	Evaluation of Image Registration Algorithms	33
2.4	Machine Learning Models in Image Analysis	34
2.4.1	Ensemble of Decision Trees (Decision Forests)	34
2.4.2	Convolutional Neural Networks	37
2.5	Applications of ML Models in Medical Imaging	39
2.5.1	Decision Forests in Medical Imaging	42
2.5.2	Convolutional Neural Networks in Medical Imaging	43
3	Multi-Modal Cardiac Image Analysis through Probabilistic Edge Maps	50
3.1	Introduction	51
3.2	Probabilistic Edge Maps (PEMs)	52
3.3	MR-US Cardiac Image Registration with PEMs	55
3.3.1	Methodology	57
3.3.2	Experiments and Results	60
3.3.3	Discussion and Conclusion	63
3.4	Multi-Atlas Segmentation of Cardiac Ultrasound Images with PEMs	63
3.4.1	Methodology	66
3.4.2	Experiments and Results	69
3.4.3	Discussion and Conclusion	72
3.5	PEM Based Respiratory Motion Correction in Cardiac Cine-MR Images	73
3.5.1	Methodology	75
3.5.2	Experiments and Results	76

3.5.3	Discussion and Conclusion	77
3.6	Other Applications of PEMs	78
3.7	Conclusion	80
4	Stratified Decision Forest: Applied to Anatomical Landmark Localisation	82
4.1	Introduction	83
4.1.1	Related work	85
4.1.2	Contributions	87
4.2	Methodology	89
4.2.1	Model Input Space	89
4.2.2	Structured Classification	90
4.2.3	Regression of Landmark Locations	91
4.2.4	Stratification of Global Characteristics	93
4.2.5	Visualization of the Stratification Splits	95
4.3	Experiments and Results	97
4.3.1	Anatomical Landmark Localisation	98
4.3.2	Cardiac MR Image Segmentation	102
4.3.3	Implementation Details	104
4.4	Discussion	105
4.5	Conclusion	107
5	Learning to Generate High-Resolution Image Representations	108
5.1	Introduction	109
5.1.1	Related work	109
5.1.2	Contributions	110
5.2	Methodology	111
5.2.1	Single Image Network	112
5.2.2	Multi-Image Network	113
5.3	Experiments and Results	114
5.4	Discussion and Conclusion	117

6	Learning Anatomical Shape Priors through Neural Networks	119
6.1	Introduction	120
6.1.1	Clinical Motivation	122
6.1.2	Contributions	124
6.2	Methodology	125
6.2.1	Medical Image Segmentation with CNN Models	126
6.2.2	Convolutional Autoencoder Model and ACNN-Seg	127
6.2.3	Medical Image Super-Resolution (SR) with CNNs	129
6.2.4	T-L Network Model and SR-ACNN	131
6.2.5	Learnt Hidden Representations	133
6.3	Applications and Experiments	134
6.3.1	Clinical Datasets	135
6.3.2	Training Details of the Proposed Model	136
6.3.3	Cardiac Cine-MR Image Segmentation	137
6.3.4	Cardiac 3D Ultrasound Image Segmentation	140
6.3.5	Cardiac MR Image Enhancement	142
6.3.6	Learnt Latent Representations and Pathology Classification	144
6.4	Discussion and Conclusion	146
6.5	Implementation Details	147
6.5.1	Network Structure	148
7	Summary and Future Work	151
7.1	Summary	151
7.1.1	Achievements	153
7.2	Limitations and Future Work	156
	List of Publications	159
	Bibliography	163

List of Tables

2.1	Applications of decision forest in medical image analysis.	41
2.2	Variants of decision forest model have been developed to solve different natural image analysis tasks	42
2.3	A list of cardiac MR and ultrasound image segmentation frameworks that utilise deep neural network architectures.	45
2.4	Single image super-resolution frameworks that make use of a trained CNN model	46
2.5	A list of other application areas of CNN models in cardiac imaging	47
3.1	3D-US cardiac image left-ventricle segmentation results	70
3.2	Accuracy of clinical index measurements - PEM multi-atlas 3D-US segmentation	71
3.3	End-diastolic volume measurement accuracy after PEM based respiratory motion correction.	77
4.1	A short summary of related work, decision forest based classification and regres- sion approaches in the literature	88
4.2	Combined landmark localisation errors for two separate datasets	99
4.3	Individual landmark localisation errors for two separate datasets	101

4.4	Estimation of global pose characteristics of the heart	102
4.5	Stratification forest parameter specifications	105
4.6	Localisation results for RV insertion points on 2D short axis slices	106
5.1	Quantitative comparison of different image upsampling methods.	115
5.2	Quantitative comparison of different multi-image super-resolution models in terms of image quality metrics	116
5.3	Cardiac MR segmentation results when super-resolution is used as pre-preprocessing	116
6.1	Cardiac MR left-ventricle segmentation results using different neural network architectures (w/wout shape priors)	138
6.2	3D-US cardiac left-ventricle segmentation results using different neural network architectures (w/wout shape priors)	141
6.3	Cardiac MR image super-resolution results using different neural network archi- tectures (w/wout shape priors)	143
6.4	Neural network architecture of the predictor used in the T-L model	148
6.5	Neural network architecture of the auto-encoder used in the T-L model	149
6.6	Neural network architecture of the baseline segmentation and super-resolution models	150

List of Figures

1.1	The anatomy of the human heart.	2
1.2	Layers of the heart wall and mitral valve of the left ventricle	3
1.3	A wiggers diagram: the electro-mechanical activity of the heart in a cardiac cycle	4
1.4	ECG synchronisation of cardiac MR image acquisition	5
1.5	Acquisition of cine MR images over multiple cardiac cycles	6
1.6	The standard cardiac cine MR imaging planes	8
1.7	Ultrasound imaging modes and image reconstruction	10
1.8	Standard transthoracic cardiac ultrasound views	11
1.9	Cine CMR stacks: low resolution in the through plane direction and respiratory motion artefacts	12
1.10	3D-US imaging artefacts in echocardiography	13
2.1	Examples of automatic cardiac MR and US image segmentation	19
2.2	Cardiac MR image segmentation with GMM-EM algorithm	21
2.3	Diagram of a standard multi-atlas label fusion segmentation method	22

2.4	Use of anatomical priors in cardiac image ventricle segmentation	24
2.5	Visualisation of multi-stack cardiac MR image super-resolution	26
2.6	Model based single image super-resolution approaches	28
2.7	Anatomical correspondences between cardiac US and MR images	31
2.8	Categorisation of image similarity metrics used in multi-modal image registration.	32
2.9	A simple decision tree structure and image categorisation	35
2.10	Visual comparison of different classification models and their decision boundaries	36
2.11	Neural connectivity comparison between convolutional and dense layers	37
2.12	Structure of a standard convolutional neural network model	38
2.13	Number of publications related to machine learning over the last 20 years	40
2.14	Number of medical publications that use a neural network model, and their distribution with respect to the problem type.	43
2.15	Visualisation of intermediate feature maps generated with a cardiac image seg- mentation CNN model	49
3.1	Probabilistic Edge Maps (PEMs): Applied to cardiac images from different modalities	52
3.2	A diagram of structured decision forest	53
3.3	A diagram of structured regression forest	58
3.4	A block diagram of the proposed multi-modal image registration method.	59
3.5	PEM - US/CT image registration results	61
3.6	PEM - US/MR image registration results	62

3.7	PEM representation of 3D-US cardiac images	65
3.8	3D-US multi-atlas cardiac image segmentation framework	67
3.9	3D-US multi-atlas cardiac image segmentation results	69
3.10	Cine-MR respiratory motion correction framework	75
3.11	Respiratory motion correction results for cardiac cine-MR image	76
3.12	An image analysis framework to compare cardiac imaging modalities	79
3.13	The differences between the US/MR imaging modalities for LV analysis	79
3.14	PEM representations for abdominal CT images	80
4.1	A comparison between standard Hough forest and stratified decision forest	84
4.2	An illustration of the stratified decision tree structure	92
4.3	Promixity analysis of decision tree training data	96
4.4	Hough vote maps and PEM representations obtained with the stratified forest . .	98
4.5	Histogram of the landmark localisation errors	100
4.6	Cardiac MR image multi-atlas segmentation results, which is initialised with the automatically localised anatomical landmarks.	103
5.1	A visual comparison of different image upsampling techniques	110
5.2	The proposed single image super resolution network model	112
5.3	The proposed Siamese multi-image super resolution network model.	113
5.4	High-Resolution image reconstruction accuracy (PSNR)	114
5.5	Qualitative comparison of multi and single input super-resolution models	117

6.1	Advantages of using anatomical shape priors in MR image super-resolution and segmentation problems.	123
6.2	Block diagram of the baseline segmentation and super-resolution CNN models. .	125
6.3	Block diagram of the proposed T-L regularisation network.	127
6.4	Training scheme of the shape regularised CNN segmentation model	128
6.5	Training scheme of the shape regularised CNN super-resolution model	132
6.6	Histogram of the learnt low dimensional latent codes (shape priors)	133
6.7	Cardiac MR segmentation results w/wout shape priors	139
6.8	3D-US cardiac image segmentation results w/wout shape priors	142
6.9	Cardiac MR image super-resolution results w/wout shape priors	144
6.10	Anatomical variations captured by the latent representations in T-L network . .	145

Chapter 1

Introduction

1.1 Clinical Background and Motivation

In this section, we give a brief introduction to the anatomy and physiology of the human heart, with a particular attention to electro-mechanical events that occur during the cardiac cycle. Later, MR imaging is discussed and we highlight the specific techniques that are used particularly in cardiac MR imaging, such as ECG gating, respiratory motion compensation, and cine imaging technique. Similarly, we give a brief summary of ultrasound imaging and its clinical role in the diagnosis of cardiovascular diseases (CVD). The section concludes with a presentation of the quantitative indices that are used in the assessment of CVD and cardiac function analysis.

1.1.1 The Anatomy and Function of the Heart

The human heart [250] pumps blood through the circulatory system supplying oxygen and nutrients to the cells of the body while removing carbon dioxide and other wastes. It has four chambers: the atria (upper) and the ventricles (lower) as shown in Figure 1.1. The atrio-ventricular septum separates the two sides of the heart. Unless there is a septal defect, the two

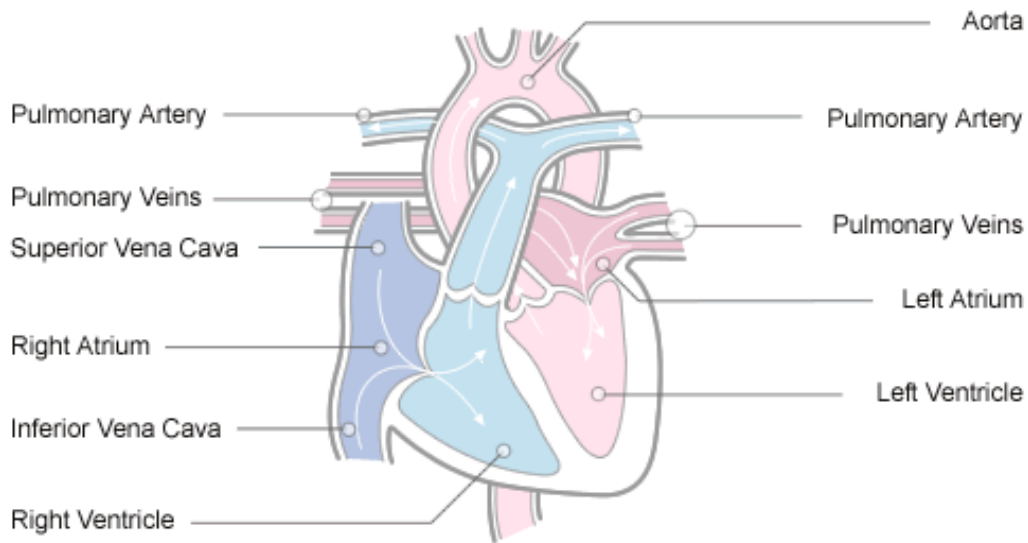


Figure 1.1: The anatomy of the human heart: coronary blood vessels, ventricles and atria. The arrows show the flow of blood through the heart¹.

sides of the heart do not directly interact with each other, but they function together.

Blood is circulated through the heart in two separate pathways [194]: the pulmonary circuit and the systemic circuit. In the former, de-oxygenated blood returns to the right side of the heart and it is pumped into the right ventricle (RV) and then to the lungs where oxygen is absorbed. At the end, oxygenated blood returns to the left atrium (LA) via the pulmonary vein. In the systemic circuit, left ventricle (LV) pumps the oxygenated blood into the aorta and arterial circulation. The bicuspid (mitral) and tricuspid atrioventricular (AV) valves separate the left and right atria, and ventricles respectively. The papillary muscles are projections of the ventricular muscles and they attach to the cusps of the AV valves (Figure 1.2). During the ventricular contraction, the papillary muscles in both ventricles tense and prevent backward flow of ventricular blood into the atrial cavities.

The ventricular wall consists of three layers: epicardium (the outermost wall layer), myocardium (the middle layer), which contains the muscle for contraction, and endocardium (the inner layer) which has contact with the blood-pool. In Figure 1.2, it can be seen that the muscle cells build up the majority of the ventricular wall, and LV wall is thicker than the RV wall since sufficient

¹The figure has been adapted from <http://www.nottingham.ac.uk/nursing/practice/resources/cardiology/function/anatomy.php>

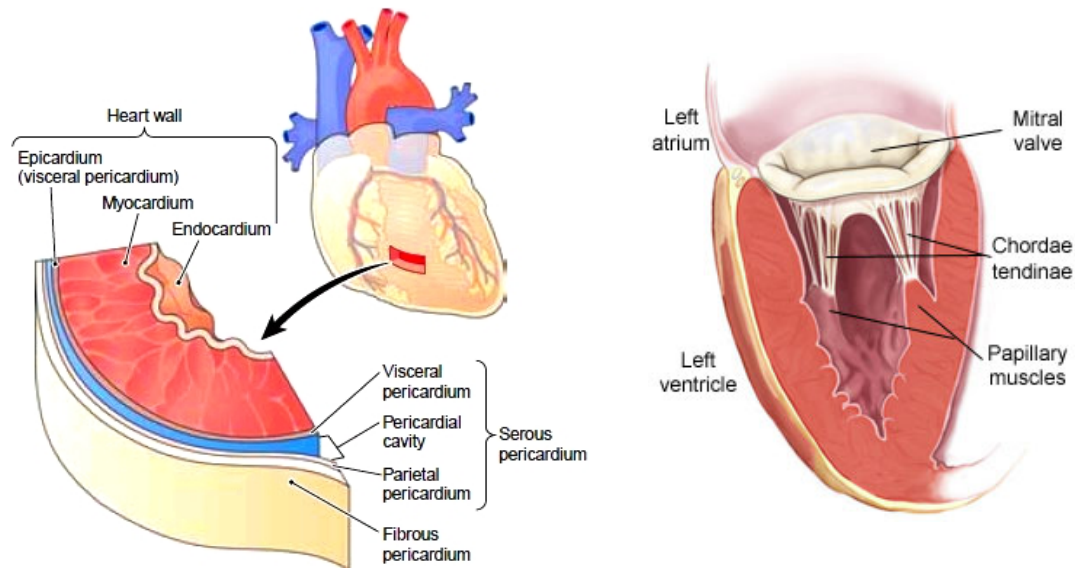


Figure 1.2: Layers of the heart wall (left) and mitral valve of the left ventricle (right)².

blood pressure is required to pump the oxygenated blood to different parts of the body.

1.1.2 The Cardiac Cycle and Electrical Activation

The cardiac cycle is a sequence of electrical and mechanical events that take place between the relaxation (diastole) and contraction (systole) phases of the heart. The ventricular diastolic stage involves the movement of blood from atria to ventricles and the ventricular systole involves the movement of blood from ventricles to the pulmonary artery and aorta. Cardiac systole is the mechanical response of the myocardial cells to an electrochemical stimulus originating from the sinoatrial (SA) node. It regulates the cardiac cycle by acting as a pacemaker. The electrical activity initiating from the SA node propagates through the fibrous skeleton of the heart (first the atrial mass and then AV node) and the resulting wave of depolarisation from top to bottom of the heart triggers the mechanical activation (*cf.* [27, 194]).

The conduction of the electrical activity through the fibrous skeleton can be seen on an electrocardiogram (ECG), as shown in Figure 1.3. The ECG recording reflects the electrophysiologic

²The figure has been adapted from <https://en.wikipedia.org/wiki/Pericardium> and <https://www.knowyourbody.net/mitral-valve.html>

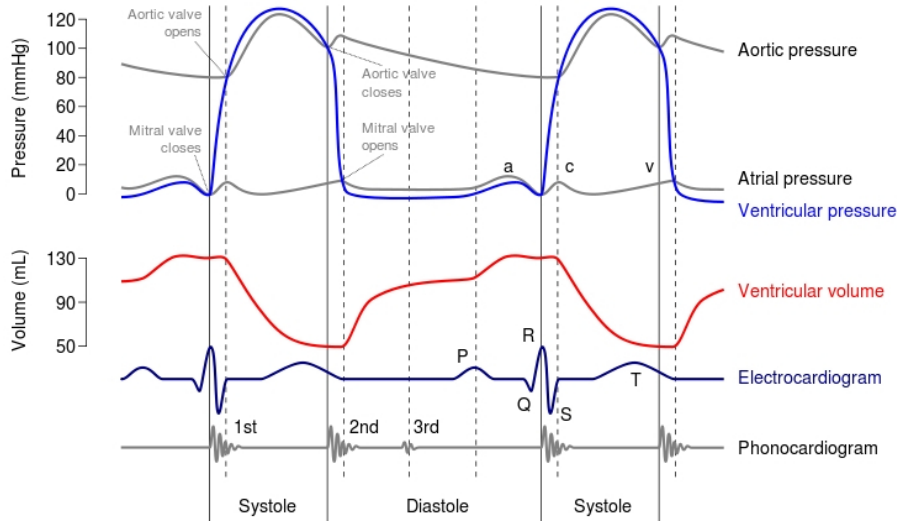


Figure 1.3: Wiggers diagram showing the electro-mechanical activity of the heart in a cardiac cycle. The figure has been adapted from https://en.wikipedia.org/wiki/Wiggers_diagram.

activity during each heartbeat and it is acquired using electrodes placed on the skin. In the same figure, the conduction at the atria is seen as the P-wave and it is followed by the PR interval that corresponds to the delay in the AV node. The spread of the electrical activity through the ventricular myocardium produces the QRS complex, and the ventricular re-polarisation (relaxation of the muscles) is seen as the T-wave. The ECG signal is also commonly used in imaging systems as a gating signal to capture images of the heart at different phases of the cardiac cycle.

1.1.3 Cardiac Magnetic Resonance (CMR) Imaging

Magnetic resonance imaging (MRI) is a non-invasive and non-ionising imaging technique that can produce tomographic images with an unmatched soft tissue contrast. Unlike X-ray computed tomography (CT), it does not use ionising radiation, but is based on the absorption and emission of energy in the radio frequency (RF) range of the electromagnetic spectrum [47]. MR images are produced based on spatial variations in the phase and frequency of the RF energy being absorbed and emitted by the anatomical tissues. Under the influence of a strong magnetic field, magnetically aligned hydrogen nuclei generate transverse magnetisation as a response to

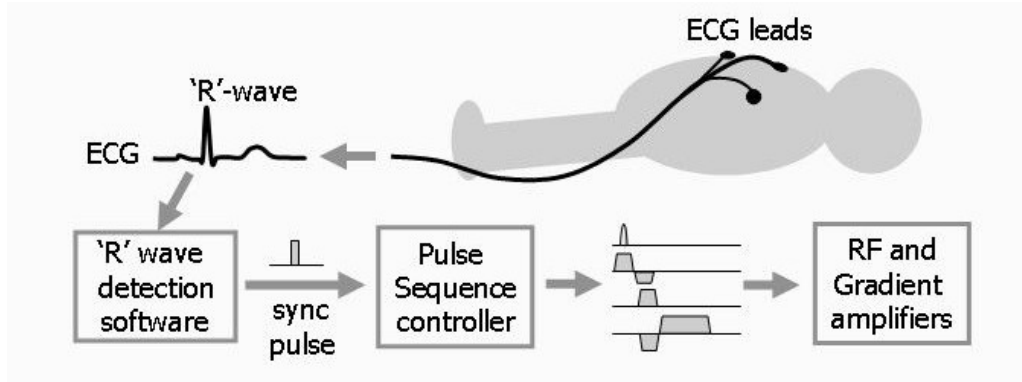


Figure 1.4: ECG synchronisation of the pulse sequences. ECG recording is used to detect the QRS complex, which triggers the pulse sequence generator to produce RF and gradient pulse waveforms. This is repeated for each cardiac cycle. Image courtesy John Ridgway [220].

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 [33].

In the context of cardiac imaging, cardiac MR (CMR) has been clinically used for evaluation of the anatomy and function of the heart chambers, valves, and major vessels. Assessment of infarct segments, myocardial wall abnormalities, LV volume measurement are some examples of its clinical use [257]. The technology has been limited in the past because of imaging difficulties associated with cardiac and respiratory motion and long acquisition time. In the recent years, CMR imaging has found more applications in cardiovascular disease [39] with advancements in spatial and temporal resolution and reduced scanning times. In particular, parallel imaging is a commonly used approach to reduce imaging time [166] and ECG gating is often necessary to reduce cardiac motion artefacts.

Electrocardiogram gating

To acquire an image of the heart unaffected by the cardiac motion, sampling of the magnetic signal needs to be done within tens of milliseconds. This situation limits the number of phase encodings (spatial resolution) and also requires the relaxation time (image quality) to be short. While such acquisition can be achieved in practice, it will come at the expense of low imaging

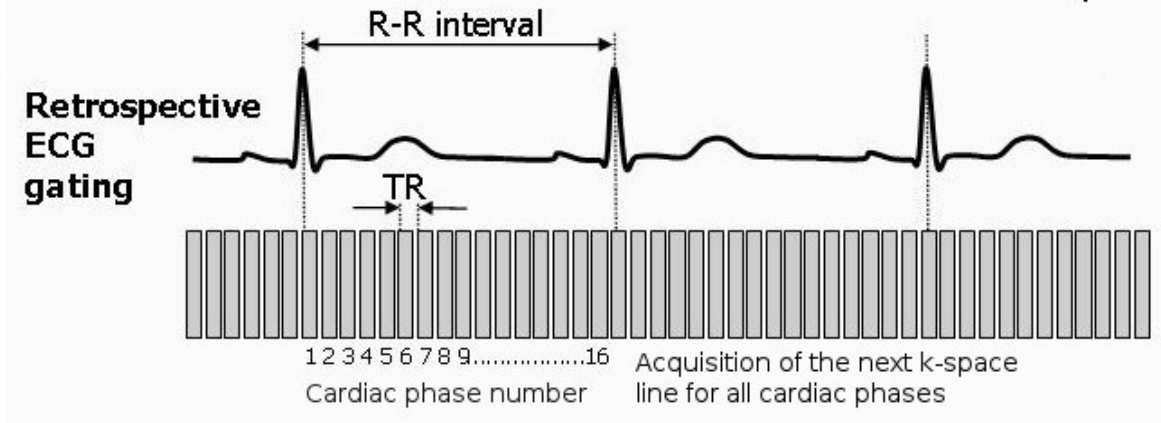


Figure 1.5: Cine MR images for all cardiac phases ($N = 16$) are acquired over multiple cardiac cycles in segments of k-space data during breath-hold. ECG gating is used to synchronise the RF pulse sequences

quality and resolution. For this reason, in standard CMR imaging protocols, k-space data sampling for each slice is performed over multiple cardiac cycles by synchronising the RF pulse sequence with respect to a fixed set of reference cardiac phases. An MR compatible ECG device is used to synchronise the MR data acquisition, which detects the R-wave and triggers the sequence controller as shown in Figure 1.4.

Respiratory motion

In conventional CMR imaging techniques, the phase encoding of each 2D image slice requires data acquisition over multiple cardiac cycles, which can result in imaging times of several minutes. This means that respiratory motion may introduce data artefacts and inconsistencies between k-space data acquired at different cardiac cycles. As a solution to this problem, there are three approaches [220]: (I) retro-respective gating of the images with signal collected from a respiratory belt, (II) single-shot fast imaging techniques, and (III) repeated breath-holding.

Cine cardiac MR imaging

Cine MR imaging visualises the temporal dynamics of the heart chambers and can be 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 image for all the cardiac phases. The images are later viewed as a movie sequence. The cine image acquisition procedure is shown in Figure 1.5. Breath-held cine MR imaging allows acquisition of k-space data in segments for each cardiac phase. In this way, the acquisition times can be shortened even further at the expense of reduced temporal resolution, which can be circumvented with echo-sharing [80].

As a pulse sequence, spoiled gradient echo (GE) has been preferred for functional imaging since very short repetition times are required. 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 flow analysis [220]. Spin echo (SE) sequence, on the other hand, is more suitable for anatomical imaging. It introduces relatively little artefacts from metal but require longer acquisition times. It plays more a secondary role in CMR imaging, but for purposes such as structural assessment of ventricular abnormalities it still proves useful [80]. Steady state free-precession (SSFP) [195] is an alternative to the GE approach, it enables relative independence of contrast from blood-flow and high speed of acquisition, which has been a limitation for very short RF repetition times [38].

Cardiac MR imaging planes

3D high-resolution cine imaging of the heart is challenging due to long acquisition times, cardiac motion, and repeated breath-holds. For this reason, stacks of thick 2D slices are acquired from different imaging planes for multiple cardiac phases. The standard cardiac cine imaging planes, shown in Figure 1.6, include short-axis (SAX), horizontal long-axis (four chamber view), and vertical long-axis (two chamber view) [93]. To identify these planes during the acquisition, first scout imaging is performed with a fast single-shot sequence, then the long-axis (LAX) planes are prescribed along a line extending from the cardiac apex to the centre of the mitral valve. Finally, the SAX plane is identified which extends perpendicular to the LAX of the heart at the level of the mid left-ventricle. Ventricular volumetric measurements are often performed on SAX

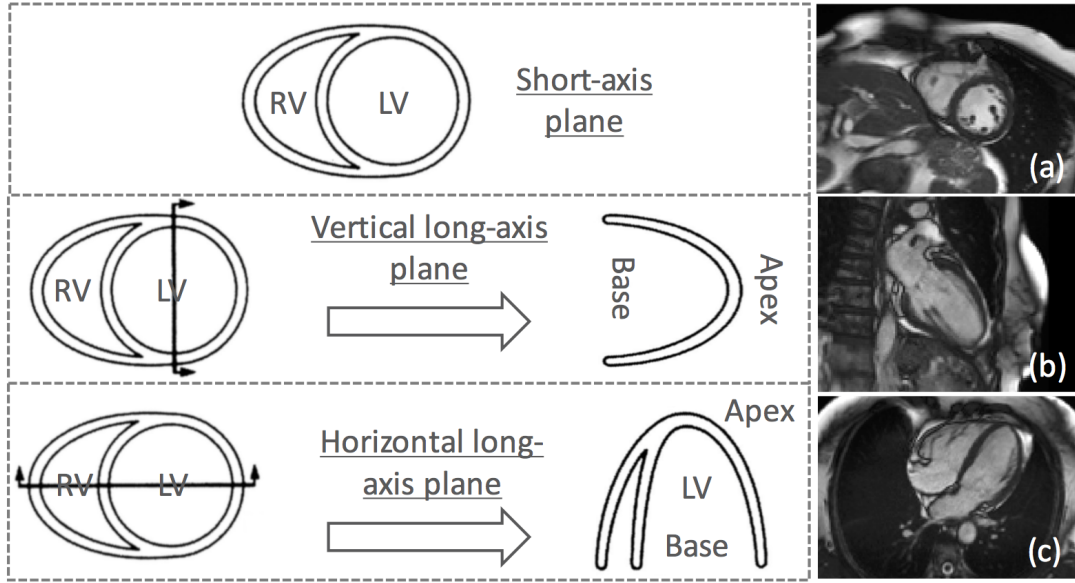


Figure 1.6: The standard cardiac cine imaging planes [93]: short-axis mid-ventricle plane (a), vertical long-axis plane (b), and horizontal long-axis plane (c). The images (b-c) are also known as two-chamber and four-chamber views respectively. The figure is adapted from [41].

stacks since it yields cross-sectional slices almost perpendicular to the myocardium boundaries. This way partial volume effects can be reduced and ventricular measurement accuracy can be improved [112]. Typically the in-plane and through plane resolutions of the SAX stacks are in between 1-2.5 mm and 8-10 mm respectively.

1.1.4 Indices of Cardiac Function

Quantitative analysis of the cardiac ventricles using imaging data starts with delineation of the endocardial and epicardial boundaries of the myocardium. Once the contours are defined for each slice in the stack of images, local and global volumetric measurements can be performed for the assessment of ventricular function and mass [112]. These measurements have clinical importance in diagnosis of cardiac related pathologies such as hypertrophy and dilated cardiomyopathy [219]. Myocardial mass (M) is a particular example. It corresponds to the weight of the heart muscle and is computed by multiplying the ventricular volume (V) calculated from the contours, with the density of the myocardium ($\rho_m = 1.05g/cm^3$) [84]. Left-ventricular (LV) myocardial mass is computed as: $M_{LV} = V_{LV} \cdot \rho_m$.

Global functional indices indicate the overall performance of cardiac ventricles in their ability to supply blood to the rest of the body [85]. They require myocardial contouring in 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: $SV = V_{ED} - V_{ES}$. Ejection fraction (EF) is the fraction of stroke volume (ejected blood) with respect to the volume of the filled heart (V_{ED}), and is defined as: $EF = SV/V_{ED}$. Another functional index is the cardiac output which 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 about which parts of the ventricle have reduced or altered contractile function. Additionally, there could be instances where global measurements fall within the healthy range while the wall motion be abnormal. For these reasons, local functional analysis of the ventricular wall is performed including wall thickening and strain analysis, which can precisely identify reversibly injured yet viable parts of the myocardium [40]. A recent survey study [85] on local cardiac wall motion provides more detailed information for interested readers.

1.1.5 Cardiac Ultrasound Imaging

Standard ultrasound (US) imaging involves a piezoelectric transducer and a processing unit for visualisation. Ultrasonic pulses generated by the transducer travel through tissues of the body and encounter interfaces and scatterers which generate echoes. These reflections occur when US wave encounters an interface with different acoustic impedance. The reflected waves are later sensed by the same transducer and converted into signal readings. The travel time and direction of the US beam is used to reconstruct the depth image along a single line as shown in Figure 1.7. A cross-sectional 2D image can be formed by transmitting pulses in multiple directions from an array using beamforming [159], and the visualised intensity values correspond to the energy of the attenuated echo signal (B-mode imaging) [115]. 2D-phased and 3D-matrix array

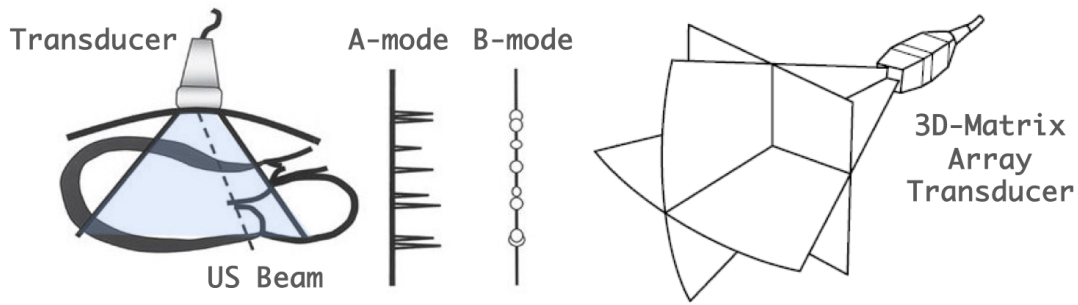


Figure 1.7: The reflected ultrasound (US) beam is captured by the transducer and the area highlighted in blue scanned the same way using multiple beams which corresponds to B-mode. A-mode shows the amplitude of the received echo for a single beam. 3D-US imaging is done with a matrix array that is composed of many piezoelectric crystals.

transducers are composed of multiple piezoelectric elements and they can automatically collect a series of 3D volumes by steering the ultrasound beam electronically [232]. This is particularly well suited for cardiac applications and typically 10-20 volumes per second are collected [115].

US cannot compete with tomographic imaging systems (*e.g.* MR, CT) in terms of detailed visualisation of soft-tissues and bony structures. However, it is capable of imaging tissue deformations real-time for surgical guidance, and its indispensable for assessment of cardiac function due to its portability, wide availability, and low cost [186]. It can provide a wealth of information [185], including the size and shape of the heart (internal chamber size quantification), estimation of LV volume, pumping capacity, regional wall motion, and the location and extent of any tissue damage. In particular, these measures are used in diagnosis and assessment of ischaemic heart disease. In cardiac US imaging (echocardiography), three standard transducer positions are used for data acquisition: parasternal, subcostal, and apical views. Figure 1.8 shows images collected from the standard trans-thoracic imaging windows. The quality of US data is often operator dependent and hence challenges for automatic analysis vary depending on the view and artefacts such as shadowing and reverberations [185]. Besides the B-mode imaging, US Doppler imaging technique provides information about the velocity and direction of blood flow in the heart, which is used in assessment of cardiac valve areas and function [136].



Figure 1.8: Standard trans-thoracic cardiac US imaging windows. Apical 4-CH and parasternal long-axis views are shown in the middle and right respectively. Image courtesy of Yao [278].

1.2 Technical Challenges in Multi-modal Cardiac Image Analysis and Objectives

Quantification of ventricular volume and mass is vital for assessment of the cardiovascular pathologies, and CMR is the reference standard imaging technique for quantitative analysis [80, 194]. There are, however, several limitations of cine CMR, such as limited through plane image resolution and respiratory motion artefacts which arise due to the need to acquire cine images over several cardiac cycles. Similarly, imaging artefacts observed in 3D-US images can hinder computer-aided analysis of the heart function. The technical limitations imposed by the imaging systems necessitate design of robust and accurate image analysis frameworks for quantitative assessment. Besides accurate predictive performance, ease of use and interpretability are essential for use of these frameworks in clinical diagnostics.

Limited through plane resolution and imaging artefacts in CMR

Cine CMR images are acquired to capture both spatial and temporal information from the heart, and long scanning times are required due to repeated breath-holds and acquisition at different cardiac phases. As a compromise between SNR and acquisition time, slices in the SAX direction are acquired with relatively large slice thickness as shown in Figure 1.9. This

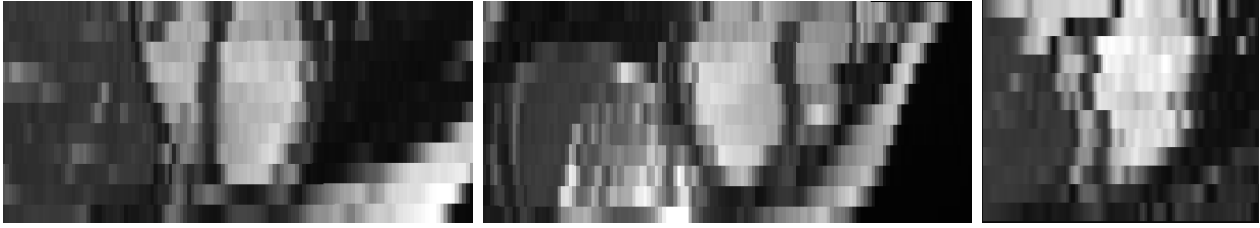


Figure 1.9: Cine CMR stacks are often acquired with relatively thick slices (8-10 mm). As slices are acquired over multiple breath-holds, there could be respiratory motion artefacts (right), which can hamper quantitative assessment when it is combined with large slice thickness.

can reduce the accuracy of the volumetric measurements (*e.g.* EF, SV) obtained with automatic segmentation tools (contouring). For this reason, most of the segmentation algorithms perform myocardium delineations for each slice independently without much considering the through-plane information [233]. There could be a few disadvantages associated with this type of 2D analysis approach: (I) partial volume effect due to low resolution (LR) can introduce error in the measurements, (II) respiratory motion artefacts observed between adjacent slices (Figure 1.9) can yield blocky disconnected contours in the through-plane direction, (III) without sufficient contextual information, segmentation of the apical and basal slices becomes a more difficult task. As a solution to these problems, research has focused on different approaches: (I) motion correction on stacks of 2D SAX images prior to segmentation, (II) sub-pixel segmentation using only the LR image data, and (III) cardiac image super-resolution.

Imaging artefacts in cardiac 3D-US images

Ultrasound (US) imaging has been preferred to be the primary imaging modality for initial assessment of the cardiac function due to its wide availability and cost-effectiveness [186]. However, quality of acquired images and accuracy of quantitative evaluations are often operator-dependent; as such, depending on the transducer position and angle, the quality of the US images can vary significantly. Common imaging artefacts [21] (Figure 1.10) that might be encountered in clinical cardiac US data are: (I) restricted field of view of the cardiac chambers (US beam profile), (II) signal drop-out caused by highly reflective anatomical structures (ribs), which can prevent imaging of the entire endocardium, (III) stationary reverberations due to

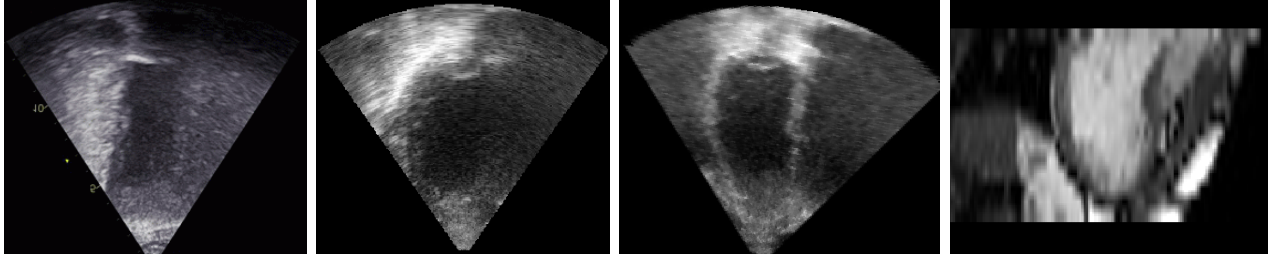


Figure 1.10: From left to right: (I) Stationary reverberations covering the anterior wall (AW) in the 2-CH image, (II) Signal drop-out of the AW in the 2-CH view due to a lung shadow distant from the probe. (III-IV) 3D-US and MR scans acquired from the same subject. Some of the structures are difficult to identify and locate in the apical ultrasound view such as the apex and right ventricle. This situation introduces a challenge in multi-modal image alignment.

multiple reflections of the ultrasound beams, which often occlude the apical part of the myocardium. These artefacts can introduce inaccuracies in cardiac ventricle segmentation, where imaging data may not be self-sufficient for precise analysis. On top of that, the low intensity contrast between the myocardium and endocardium can make the segmentation problem even more challenging, as shown on the left in Figure 1.10. Image analysis frameworks have tried to address these challenges through use of anatomical shape models [203, 252] and techniques such as stitching or compounding multiple acquisitions [278].

Identification of correspondences in multi-modal imaging data

Multi-modal imaging can provide complementary information that may not be extracted via single modality alone. For instance, 3D-US data provide useful guidance in surgical procedures through visualisation of the papillary muscles and high resolution temporal information [186] which are not usually captured by CMR images. More importantly, CMR is not compatible with many surgical devices and thus difficult to use during interventions. The fusion of the multi-modal data, however, often requires spatial alignment of the multi-modal images, thus correspondences between US and MR images are required to be identified at least on a sparse level. This is a challenging task because some anatomical structures are not visualised in both modalities at the same time and their appearances can be significantly different as shown on the right in Figure 1.10. Moreover, manual delineations of the endocardial border in both

modalities often do not show agreement with each other as has been demonstrated in clinical studies [170, 241]. For this reason, the volumetric measurements obtained from 3D-US and CMR often show a systematic difference. In conclusion, the multi-modal ultrasound image registration is a challenging and topic of on-going research.

User interaction requirement in semi-automatic segmentation methods

Besides the challenges imposed by the imaging systems, usability of image processing techniques is another aspect that needs to be considered. In particular, semi-automatic segmentation tools (multi-atlas, deformable models) often require guidance by an operator to perform the delineation task. In particular, operator input is often required to initialise these algorithms by identification of a particular set of landmark points in images. However, this can be a limitation for deployment of these algorithms in processing of large imaging cohorts such as UK Biobank [206], where thousands of images are required to be analysed for population studies. To address this limitation, many research efforts have focused on development of automatic landmark localisation techniques, which needs to ensure reliable and accurate segmentation accuracy while automating the analysis frameworks.

1.3 Thesis Contributions

The following sections present the contributions of this thesis in analysis of cardiac multi-modal imaging data and also summarise how the aforementioned challenges are addressed.

Edge-maps for multi-modal image analysis

We propose a novel image representation technique which we term probabilistic edge-maps (PEMs). PEM is an intermediate feature space where multi-modal image data is first projected into before analysis. Subsequent image analysis benefits from this feature space as it is

organ specific and highlights only the structural information of the anatomy, which is useful particularly in US image analysis. We demonstrate its applicability for multi-modal cardiac image registration, multi-atlas US image segmentation, and cardiac cine MR respiratory motion correction problems. The proposed approach differs from existing image feature representation types by utilising a trained decision forest (DF) to perform the feature mapping, which provides supervision and robustness against imaging artefacts.

Automatic anatomical landmark localisation in CMR images

We demonstrate that the anatomical landmark localisation problem in cardiac imaging can be solved with a structured DF based regression model in a very robust and accurate way. The proposed approach is utilised in a multi-atlas segmentation framework as a preprocessing technique to obtain a fully-automatic segmentation pipeline. From a more technical point of view, a new decision tree model is presented, which can exploit both local and global level image annotations such as subject-specific meta information, including size and scale of the target anatomy.

Cardiac MR image super-resolution with convolutional neural networks

In cardiac imaging, image super-resolution (SR) can be used to reduce the quantitative measurement errors that arise due to thick slice acquisitions. In that regard, we propose one of the first approaches that employs a convolutional neural network (CNN) model for the cardiac image SR problem and demonstrate the performance improvement in terms of accuracy and inference time by comparing it to state-of-the-art SR approaches. The presented SR framework fuses the information from multiple input images that are acquired from different imaging planes (*e.g.* SAX, LAX). More importantly, we show that the accuracy of subsequent image analysis (*e.g.* segmentation, myocardium tracking) can be improved when the input images are preprocessed with the proposed SR technique.

Learning anatomical shape priors with convolutional neural networks

The current state-of-the-art medical image segmentation and SR frameworks are based on CNN models; however, commonly used training objectives do not explicitly supervise these models to learn the global structure of the target space, such as label shape and topology. This could create a potential problem when image data is not self-sufficient for the analysis due to imaging artefacts, where anatomical priors are required for the analysis. In that regard, we explore new model training objectives that could make use of manual annotations more effectively and thus provide better model supervision. A novel training strategy is proposed to incorporate anatomical priors into a standard CNN model. With this approach the trained models are guided to respect the underlying anatomical constraints in their predictions. For instance, SR is an ill-posed problem and there could be multiple valid inverse HR solutions for a given LR image. In our solution, the proposed SR model will choose the anatomically most accurate solution among all the possible options.

Moreover, we demonstrate the advantage of end-to-end network training to circumvent traditional image pre- and post-processing techniques such as motion-correction, super-resolution, and conditional random fields. We further show that application driven model learning can remove the necessity of intermediate image analysis steps in commonly used frameworks. In other words, the models can be supervised to learn the direct relationship between the image data and target objective. For instance, 3D high resolution segmentations are obtained directly from 2D-SAX stack data without requiring any respiratory motion correction or super-resolution.

1.4 Thesis Overview

In the following chapter, background information regarding the cardiac image analysis problems and state-of-the-art methods are reviewed with a particular focus on machine learning. In particular, we discuss some of the proposed methods that involve a DF or CNN for cardiac image

segmentation, multi-modal registration and SR problems. Chapter 3 introduces the probabilistic edge-map (PEM) representation for multi-modal image analysis. Results are presented for cardiac US multi-modal image registration and segmentation problems, and compared against other commonly used approaches. In Chapter 4, the PEM framework is extended to anatomical landmark localisation by introducing the structured Hough Forest model. Its applicability in a multi-atlas segmentation framework as an initialisation method is demonstrated with the experiments. The idea of CNN based cardiac image SR is explored in Chapter 5 and its benefits on subsequent image analysis is shown with motion tracking and segmentation experiments. In Chapter 6, a novel strategy is proposed to supervise segmentation and SR CNN models with an anatomical shape model, which encourages CNN models to make anatomically more meaningful predictions. Finally, in the last chapter, we summarise the work presented in this thesis and discuss future work.

Chapter 2

Review of Machine Learning Methods in Cardiac Image Analysis

Hardware limitations imposed by the medical imaging systems can hinder diagnosis and treatment of cardiac related pathologies. To facilitate these procedures and provide better clinical outcome, computer-aided image analysis frameworks have been developed in the last two decades. This chapter gives a brief overview of these analysis frameworks and their applications in cardiac imaging for segmentation, image enhancement, and multi-modal image alignment. These computational methods share common objectives: time-efficiency, objective quantitative assessment, enhancement and analysis of multi-modal imaging data. In this survey we focus on how image representation learning can contribute to these objectives and increase the accuracy and robustness of the applied techniques. In particular, our main focus is on anatomical priors in medical image segmentation and also on intermediate feature spaces used in multi-modal image registration and super-resolution problems.

Additionally, the chapter discusses machine learning (ML) based analysis techniques which have recently gained interest in the medical imaging community with the advent of annotated datasets. In that regard, we present two commonly used models and discuss their aspects in terms of design flexibility, model parametrisation, and practical aspects (e.g. computation

requirements). Later, an overview of ML applications in cardiac imaging is provided, which include other application areas as well, such as landmark localisation and image quality control.

2.1 Cardiac MR and Ultrasound Image Segmentation

Cardiac MR (CMR) imaging is of vital importance in clinical cardiology due to its non-ionising and high tissue contrast visualisation of heart chambers and vessels. More importantly, clinical indices derived from CMR images help in the assessment of cardiac function and anatomy, through which clinicians can diagnose cardiovascular diseases (CVDs). Stroke volume (SV), ejection fraction (EF), and wall thickness measurements are obtained in standard practice for functional assessment of the left ventricle [18], and they are useful indicators of non-ischaemic cardiomyopathies such as dilated LV and hypertrophy [219]. In clinical practice, computation of

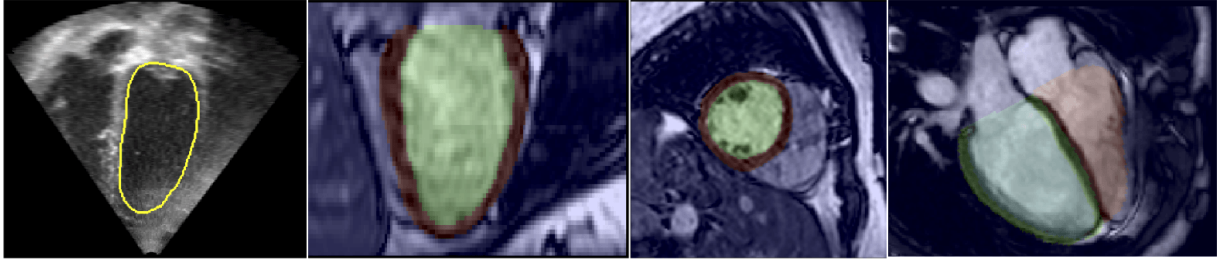


Figure 2.1: Examples of automatic cardiac MR and US image segmentation results obtained with a convolutional neural network. Cardiac short-axis (second from left) and long-axis (right-most) MR images are partitioned into semantic parts: LV myocardium, LV blood-pool, and background. The number of pixels labelled into blood-pool and myocardium classes is calculated to obtain clinical measurements such as ejection fraction and ventricular mass.

these measurements or biomarkers requires manual or automatic delineation of ventricle boundaries (e.g. endo- and epi-cardium). Manual contouring is a tedious, operator-dependent, and time-consuming task; for these reasons, researchers have studied semi and automatic ventricle segmentation approaches to obtain consistent and accurate delineations from SAX and LAX images as shown in Figure 2.1. As presented in the recent survey [205], CMR segmentation approaches can be categorised into two groups as image-driven and model-driven methods. The approaches in the latter category make use of appearance, context and anatomical knowledge

learnt from expert annotations to classify image pixels. In this way, the segmentation becomes less sensitive to low-contrast, intensity noise, and acquisition artefacts. Image-driven methods, on the other hand, are driven by minimising a pre-defined energy function that is composed of image intensity, contour interaction, and label smoothness energy terms.

2.1.1 Energy Minimisation Methods

The early work [163] on cardiac image segmentation used thresholding to identify object boundaries by clustering each individual pixel based on a fixed intensity threshold, which is often sensitive to intensity variations and results can be spatially inconsistent. Region-growing techniques [142] operate based on pre-selected seed points, and anatomical structures are segmented by propagating the seed labels to neighbouring image pixels based on their intensity values. Over-segmentation and leakage of ventricle boundaries have been reported [205] as the main limitation of these approaches. Active-contour and level-set segmentation techniques [202], on the other hand, do not approach the problem as pixel by pixel classification but contour-fitting via iterative energy minimisation. During the optimisation of the objective function, the topology and smoothness of the curve is regularised while the gradients in the image intensities guide the fitting procedure.

In some cases [46], shape priors are used in the energy function as well. However, these techniques mostly rely on accurate contour initialisation. For example, statistical models (e.g. shape PCA [52]) have been used to model the shape of the ventricles and they are used in image segmentation algorithms to enforce anatomically plausible delineations. In these approaches, optimisation is done with respect to a set of learned variation modes that characterise the distribution of surface points describing the organ shapes. Particular examples are active shape models (ASM) [143] and appearance models (AAM) [180] which require identification of landmark correspondences between target and training images. As highlighted in [205], these models can be computationally very demanding during the surface fitting procedure and may not always generalise well to previously unseen shapes. Graph-cut techniques [28] have also

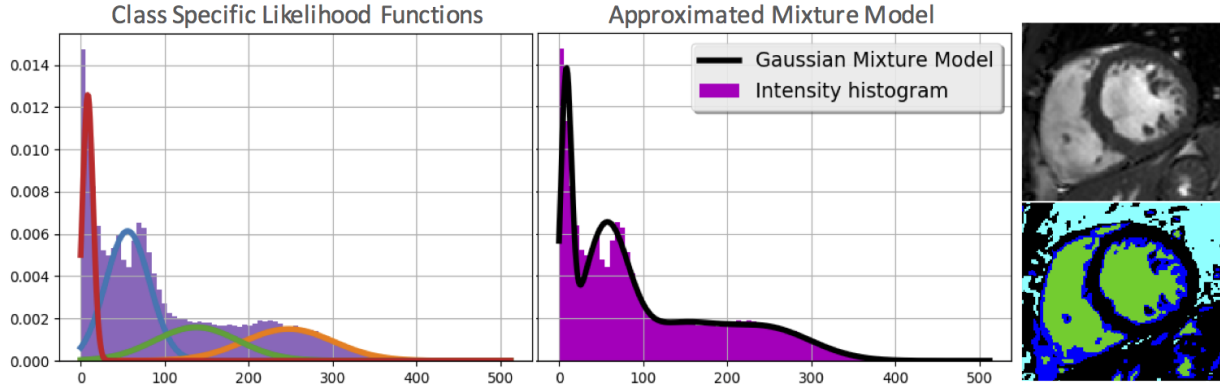


Figure 2.2: Cardiac MR image segmentation with a Gaussian mixture model using expectation-maximisation algorithm. The intensity histogram is modelled with a weighted normal distributions shown in the middle, and the class-likelihood distributions are shown on the left figure.

been widely used in both cardiac MR [169] and US [215] segmentation problems. It uses the max-flow algorithm [200] to find the minimum cut on a graph, which yield optimal solutions in polynomial time. In that regard, it has the advantage of being fast and give globally optimal results. Also, it is not sensitive to initialisation, while deformable model-based approaches are sensitive to initialisation. Extensions to the original algorithm were proposed in the later years, including computationally efficient 3D+T CMR segmentation [152] and shape prior based energy formulations [168].

2.1.2 Gaussian Mixture Models

Gaussian mixture modelling (GMM) of image intensities has been proposed in [121] as another way to segment cardiac LV. In its simplest form, this approach classifies pixels based on a fitted intensity distribution that is modelled as a weighted sum normal distributions as shown in Figure 2.2. The likelihood optimisation is performed using the expectation maximisation (EM) algorithm [64]. In follow-up work [155], a spatially consistent GMM is proposed to ensure output label consistency, which is achieved with a Markov Random Field (MRF). Moreover, spatially warped 3D+T probabilistic cardiac atlases are used in the same algorithm to guide the EM optimisation, where the atlas prior information is used to estimate the initial GMM parameters. Similarly a probabilistic atlas is used in [291] to initialise the GMM-EM framework; however,

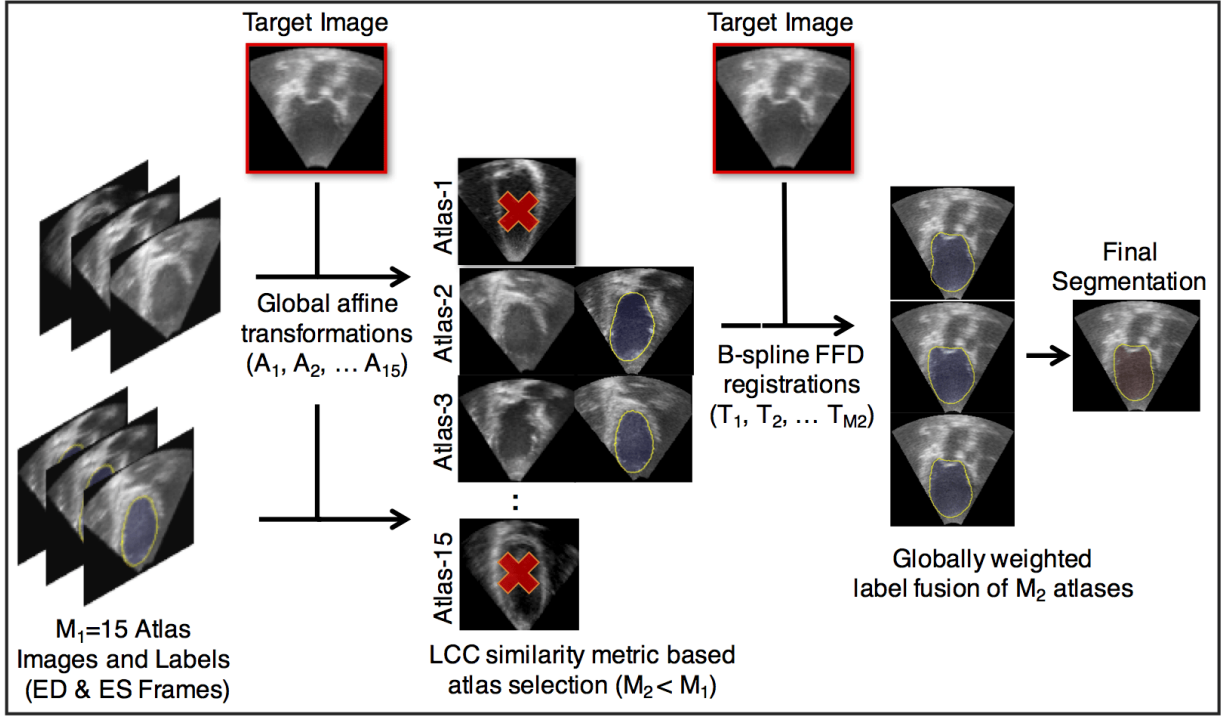


Figure 2.3: Diagram of a standard multi-atlas label fusion segmentation method. The atlases shown on the left are spatially warped to the target image (in red) using affine (A) and deformable (T) transformation models. The computed spatial mappings are later applied to atlas LV segmentation maps and fused together to obtain the LV delineation of the target image.

the authors have proposed a multi-variate mixture model to perform a joint segmentation of multi-sequence CMR images. In [244], Shi et al. have merged the EM and graph-cut optimisation techniques in the same framework to segment infarct regions on the myocardium. In a recent work [35], a joint CMR image reconstruction and GMM-EM segmentation approach has been proposed. In this method, the compressed sensing energy optimisation is guided by a mixture model so that the reconstructed intensity values can be modelled more accurately for the segmentation task.

2.1.3 Multi-Atlas Segmentation Methods

Multi-atlas label fusion (MALF) technique is another segmentation approach that has been used to delineate ventricles in cardiac MR [11, 293] and 3D-US [192] images. As shown in Figure 2.3, manually annotated atlases are warped to target image using affine and deformable

transformation models. The computed spatial transformations are applied to manual segmentations and fused together to generate the final LV segmentation. MALF requires many atlas to target image registrations, and the performance of the algorithm mostly depends on the accuracy and robustness of the atlas to target image registrations. MALF approaches possess two main features that make them reliable segmentation tools: (I) segmentation label consistency and object shape are ensured in label predictions since the atlas framework operates directly on the manual annotations, and (II) large number of training data is fused in an effective way via label fusion algorithm so that the ensemble of all warped annotations result in more accurate delineations. However, performing multiple registrations can be computationally expensive and in some cases anatomical landmarks might be required to obtain an initial alignment between target and atlas images. In comparison to other approaches, multi-atlas segmentation methods have consistently performed well and achieved state-of-the-art results in RV segmentation benchmarks [207].

It is also important to note that over the last two years, more accurate and computationally faster approaches (e.g. convolutional neural networks) have been proposed and benchmarked in cardiac LV segmentation challenges [237]. Moreover, estimation of clinical indices may not always require an image segmentation algorithm. In [288], a machine learning based framework is proposed to directly estimate EF and LVM values. As there have been different ways to measure LV volume, such as Simpson’s rule [111], a machine learning model can potentially exploit an optimal way to learn the relationship between image data and volumetric measurements in an end-to-end manner without requiring accurate ventricle delineations. Image segmentation approaches that involve more advanced machine learning models are discussed in Section 2.5.2 and a short comparison to traditional methods is provided.

2.1.4 Anatomical Priors in Cardiac Segmentation

In cardiac MR segmentation, ventricle boundaries do not necessarily align with intensity edges. For instance, trabeculae in LV blood pool are commonly included in LV endocardial segmenta-

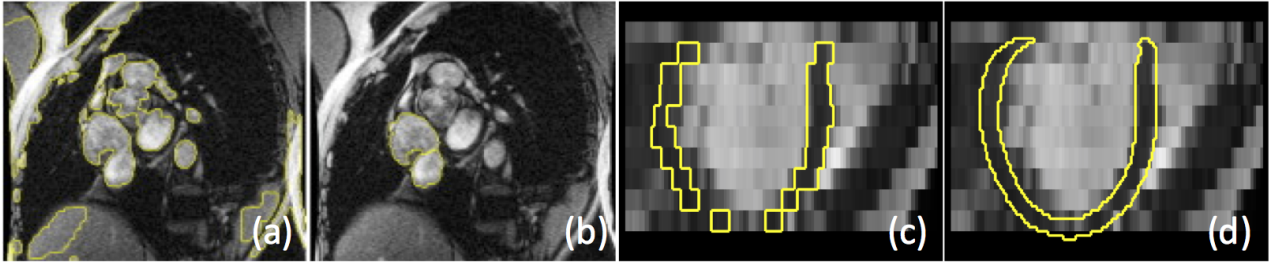


Figure 2.4: Use of anatomical priors in cardiac image left-ventricle segmentation. Topological prior is used in [281] to obtain a connected single segmentation object (a-b). Similarly, statistical shape model is used to regularise the CMR SAX image segmentation output (c-d) [86].

tions [201] although there is a large intensity contrast between the blood pool and trabeculae. For this reason, image contrast alone may not always be a reliable source of information for automatic delineation tools. Similar situations can be observed in echocardiographic images; the contour of the myocardium may be incomplete simply because it is out of the region of the transducer coverage. Additionally, noise in ultrasound images combined with signal drop-out artefacts can lead to higher prediction error in segmentation and registration methods. A more detailed discussion on these limitations (e.g. motion artefacts) is provided in Section 1.2. For these reasons, in image-driven segmentation approaches, anatomical information may need to be taken into account in design of the energy function. This can be achieved by including additional constraints in the energy function, which are related to the cardiac structure, shape statistics, or physical properties of the heart. Tavakoli et al. [261] have provided a survey study on the use of anatomical priors in cardiac image analysis methods tackling segmentation and registration problems. In this section, we briefly discuss about the prior-based methods that are relevant to the scope of this thesis. In particular, we are interested in priors that are based on a statistical model, and also the deterministic ones that are used as a regularisation term in optimisation such as adjacency and edge polarity.

As an example to the first category, PCA-based statistical shape models [52] (ASM, AAM) have been commonly used to refine the result of level-sets [265] and active contour segmentation [82] algorithms. Similarly, these models have been employed in [212] to parametrise deformation fields in non-rigid image registration. In both cases, the trained generative models are used to compute shape distances that guides algorithms to generate anatomically more plausible results.

In other segmentation frameworks, anatomical priors have been encoded as a regularisation term without requiring a statistical model or training dataset. Some common examples used in cardiac imaging can be summarised as follows: (I) Thickness prior: the radial distance between endocardium and epicardium labels has been utilised as a constraint in a level-set US segmentation framework [65]. (II) Topological prior: many anatomical objects in medical images possess a specific type of topology. For instance, LV endocardium segmentation surface is expected to be a connected object without any holes. The topology of segmentation maps can be preserved by enforcing this constraint [281] (Figure 2.4). (III) Motion prior: Segmentation and tracking have a variety of applications in both cardiac MR and US imaging. In [74], motion prior is used to constrain the temporal evolution of the level-sets segmentation algorithm.

Biophysical models belong to another category of prior information that has been used in cardiac imaging. This group of techniques relies on the mechanical, physical, or physiological properties of the heart such as finite element models [277], incompressibility of the myocardium [24], and physical or elasticity laws as constraints for segmentation and motion estimation of cardiac images. Multi-atlas techniques [11, 119] can utilise similar anatomical constraints (e.g. adjacency and shape) in an implicit manner by propagating the manually annotated labels into the target image space, which has been shown in Figure 2.3. Similarly, machine learning based image registration and segmentation approaches can integrate the prior knowledge into their framework implicitly through learning from external annotated datasets.

2.2 Super-Resolution in Magnetic Resonance Images

The sensor resolution of a medical imaging system is determined according to the physical constraints which limit the data sampling frequency and signal-to-noise ratio of the acquired data. In MR imaging, the problem arises from the need for long signal recovery between excitations to enable the operation of the spin-echo mechanism that provides tissue contrast. In addition to these limitations, acquisition time is another factor determining the resolution

of acquired images as in the case of cardiac MR imaging, where repeated breath-holds and temporal data acquisition are required. For these reasons, the acquisition of high resolution 3D+T cardiac images is very difficult to perform within clinically acceptable time frames. On

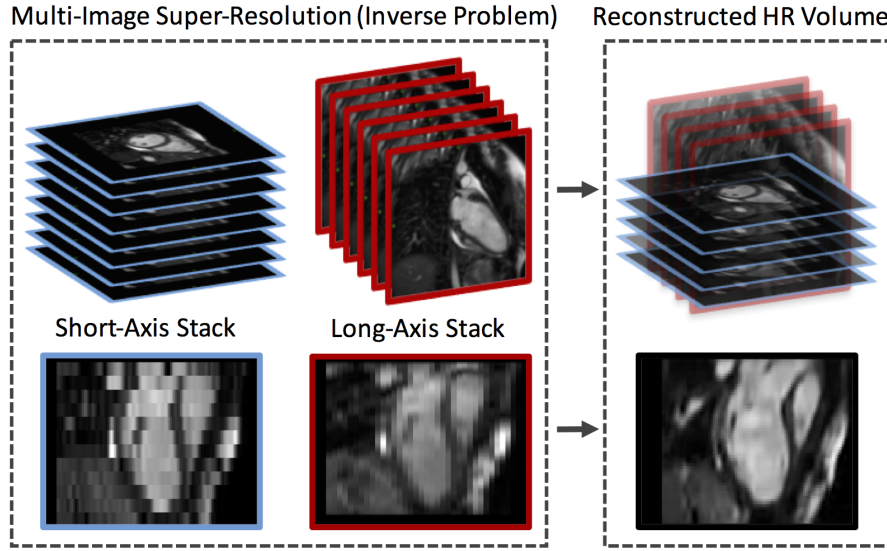


Figure 2.5: Stack of 2D cardiac MR slices acquired along the short- (SAX) and long-axis (LAX) planes. Information from multiple stacks can be merged into a single high resolution volume (shown on the right) through an image super-resolution algorithm [188]

the other hand, high resolution (HR) 3D imaging is the key to more accurate understanding of the anatomy and function of the human body. Therefore, in clinical practice [172] multiple low resolution (LR) cardiac MR images are acquired from different imaging planes (e.g. SAX and LAX stacks) and clinical assessment is performed through analysing the anatomy from these LR images as shown in Figure 2.5.

In the context of MR imaging, a common approach [103] assumes that LR images ($\mathbf{x}$) can be modelled as a weighted average of the corresponding HR image ($\mathbf{y}$) plus a noise signal ($\boldsymbol{\eta}$). In cardiac imaging, due to respiratory and heart motion, a spatial transformation operator ($\mathbf{M}$) is included in this forward model which can be formulated as $\mathbf{x} = \mathbf{D}\mathbf{B}\mathbf{S}\mathbf{M}\mathbf{y} + \boldsymbol{\eta}$. Here $\mathbf{S}$ and $\mathbf{B}$ correspond to slice selection and blurring operators respectively. The blurring kernel is associated with the slice profile of the scanner and usually approximated by a Gaussian function [103]. The last operator $\mathbf{D}$ corresponds to image decimation.

In the interest of providing more detailed visualisation, some studies [23, 243] have proposed

to tackle these hardware limitations through inverse solution techniques such as image super-resolution (SR). In standard SR approaches, a single image with a high pixel resolution is synthesised by combining independent LR acquisitions of the same organ. In other words, combination of several sources provides more information than each of the individual sources. Besides the improved image quality and resolution, SR can provide better guidance for subsequent computer-aided image analysis techniques such as image segmentation. In [228], the authors demonstrate that precision of brain tissue segmentation can be improved by pre-processing the input low-resolution MR image data with a SR algorithm. Similarly, other studies have shown that SR can be a useful image pre-processing technique in fetal MR [134] and PET imaging [126] for better visualisation of the brain tissues.

Image SR is usually an ill-posed inverse-problem that requires the approximation of the HR image data from the given LR image(s) [204]. SR approaches can be characterised in terms of two main properties: (I) whether they operate on multiple LR image data acquired from different directions or require only single LR image data, and (II) whether they operate based on a learned inverse model or relies solely on the given image data. Traditional variational approaches (e.g. total variation [8]) solve the inverse problem iteratively through energy minimisation. On the other hand, model-driven approaches [2, 23], in particular in single image SR, require a generative or regression model to approximate the inverse function and are learnt from an external dataset containing LR-HR image pairs. In exceptional cases, single image SR techniques may not require external training data as in the case of non-local means based image upsampling methods proposed in [171, 226].

2.2.1 Variational Inverse Methods

The iterative back-projection algorithm (IBP) [105] is a traditional multi-image SR approach that attempts to solve the inverse problem by making an initial guess on the high-resolution image, which is later iteratively updated by back-projecting the error signal originating from the synthetically generated LR images and original input data. To obtain anatomically plausible

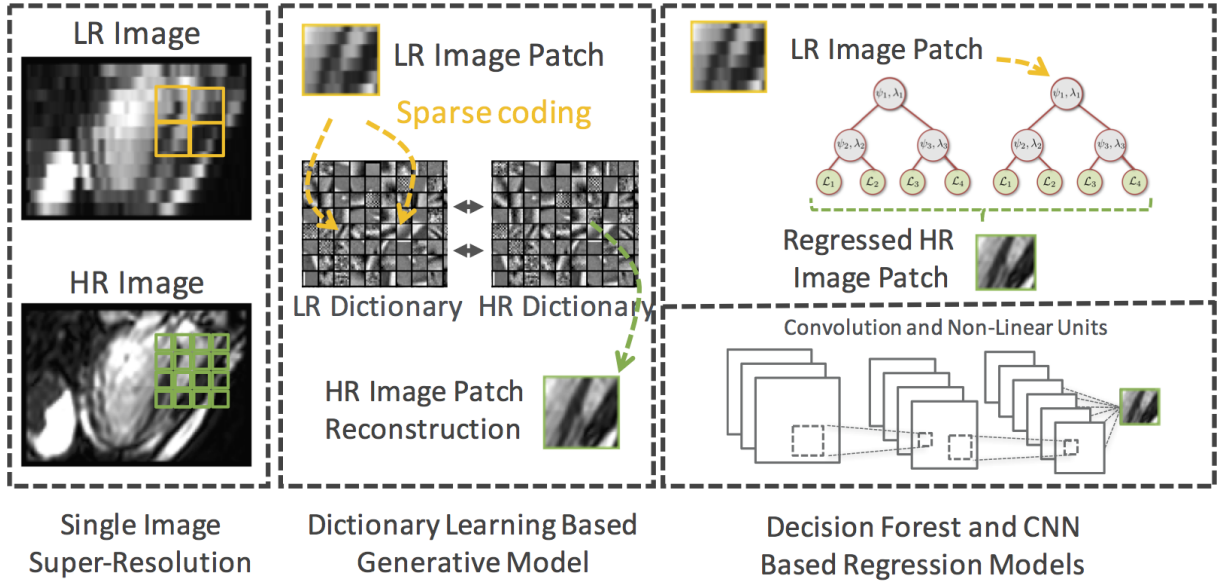


Figure 2.6: Model based single image super-resolution approaches: coupled low- (LR) and high-resolution (HR) dictionaries are used to synthesise HR image patches [23], similarly regression models can be trained solve this inverse mapping problem [2, 70].

results, the standard IBP energy function is modified by including regularisation terms such as cross-image self-similarity [208] or total-variation terms. Similarly, Odille et al. [188] proposed an energy minimisation based SR approach that performs joint reconstruction and motion compensation for cardiac 2D image stacks, which includes denoising with Beltrami regularisation. Energy minimisation based SR reconstruction techniques require a good initial approximation of the solution for both the HR volume and also for the motion model [188]; Otherwise optimisation techniques such as Gauss-Newton [25] are not guaranteed to converge to reliable SR results. More importantly, multiple stack input data might not always be available and is not acquired in standard imaging protocols. As discussed in Section 1.1.3, cardiac images acquired along the LAX plane do not necessarily always cover the whole heart.

2.2.2 Regression Models for Image Super-Resolution

In general, single image super-resolution techniques are more flexible and have more application areas since they do not require multiple data acquisitions from different views. However, external training data is often required to learn a generative SR model. Bhatia et al. [23]

proposed a coupled dictionary learning approach to model the relationship between low- and high-resolution image patches collected from cardiac MR data. Similarly, in [243] a multi-atlas technique is used to transfer high-resolution information from atlases to a low-resolution target image grid. Although multi-atlas techniques do not have an explicit parametric model, since they make use of an external dataset, they are included in this review. As shown in Figure 2.6, a regression model can be learnt to perform the mapping from LR to HR patches. In [2], a regression forest based approach is proposed to upsample diffusion tensor images. In general, these approaches differ from each other in terms of the way they model the inverse mapping between low to high resolution image patches.

In the recent years, machine learning approaches in particular CNNs have been shown to be capable of learning inverse functions very effectively, and the authors in [259] have shown their applicability in diffusion MR image SR problem. However, the presented experiments do not show whether subsequent image analysis (e.g. pathology classification) does actually benefit from SR preprocessing or not. Since HR pixel values are generated as a result of an optimisation process, resultant intensity values may not be considered as reliable as the ones in the original images which are directly generated by the reconstruction methods. In contrast, by showing the benefits of SR for subsequent analysis, one can provide a scientific evidence for its use as a preprocessing technique in analysis frameworks.

2.3 Multi-Modal Cardiac Image Registration

Imaging techniques play an essential role in the evaluation of cardiac related pathologies, interventional treatment guidance, and prognosis throughout the patient's follow-up. Integration of several imaging techniques provides multi-modal data that contain both functional and anatomical information [222]. For instance, in image guided cardiac surgery, overlaying 3D MR/CT images on real-time fluoroscopic or ultrasound data enables tissue characterization and precise localisation of instruments and catheters with respect to the underlying anatomy [96]. Similarly,

image fusion of CT angiography with cardiac perfusion MRI has facilitated detailed assessment of coronary artery disease [251]. Image fusion, which is also referred as image registration, is the task of spatially aligning and visualising different aspects of human anatomy simultaneously. This can be performed by means of a dedicated combined imaging system (e.g. PET/MRI, PET/CT) or through image registration techniques. Despite extensive research, these image registration algorithms have not been widely adopted in clinical practice due to two main reasons: (I) most registration algorithms still require manual interaction for initialisation, and (II) anatomical knowledge is needed to be integrated into algorithms to understand and match correspondences between different imaging modalities. The latter can be considered as the main challenge in multi-modal image registration. In general, image registration can be performed between two or more images. In the current context, we only consider methods that involve two images. A registration algorithm typically minimises an energy function that is composed of dissimilarity ($\mathcal{S}$) and regularisation ($\mathcal{R}$) terms:

$$\arg \min_T \mathcal{S}(I, J \circ T) + \mathcal{R}(T) \quad (2.1)$$

Optimisation techniques allow us to minimise the energy function and find an optimal spatial transformation (T) that maps a source image (J) to a target image (I). In summary, an image registration algorithm mainly consists of three main building blocks: A transformation model, an optimisation strategy, and an objective function measuring the image similarity.

2.3.1 Transformation Models and Optimisation Techniques

Optimisation techniques can be roughly grouped into two categories: continuous and discrete methods. In the latter case, registration is expressed as a discrete labelling problem over a MRF [147]. These approaches have been shown to be more robust against local-minima problem due to non-convex nature of the optimisation task [79] and they are inherently gradient-free. In addition to that, major speed improvements can be achieved as parallel architectures can be

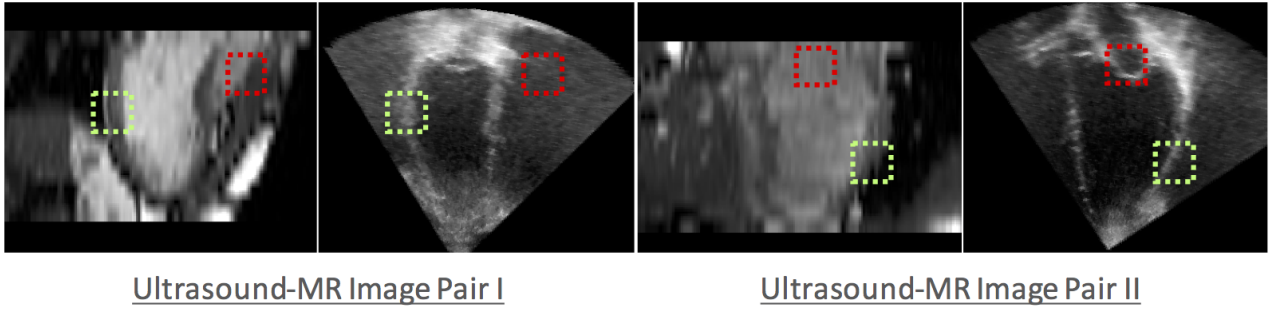


Figure 2.7: Two pairs of cardiac 3D-US and MR images, acquired from same subject, are shown. Identifying the anatomical correspondences between the two modalities is not an easy task due to spatially varying intensity statistics. Additionally, some anatomical structures might not be visible in ultrasound due to limited field-of-view and signal drop-out (shown in red). For instance, papillary muscles (#2) and right ventricle (#1) are not visible in the shown US images, which needs to be taken into account in design of similarity metrics or optimisation.

used in several discrete algorithms to perform non-sequential tasks. A more detailed comparison on registration optimisation techniques can be found in [247].

Transformation models, which define the warping function between source and target images, can be categorised into two subgroups as linear (rigid/affine) and deformable. The latter can be further divided into two subcategories as physical model and interpolation based deformation models. Common examples for the former are elastic body [12], fluid flow [48], and flow of diffeomorphisms models [16], which regulate displacement fields based on some underlying physical phenomena such as bending energy. The interpolation based deformation models use smooth basis functions with local or global spatial support such as: free-form deformations (FFD) [227], radial basis functions [26], and locally affine models [110]. Further details about deformable models can be found in the recent survey paper [247].

2.3.2 Image Similarity Criteria in Multi-Modal Image Registration

In multi-modal image registration, the formulation of similarity metric has more importance and is more challenging compared to mono-modal registration due to a few reasons: (I) Some anatomical structures may not always be visible in both modalities (e.g. US and MRI) due to physical properties of the imaging systems. (II) Anatomical tissues might be visualised with

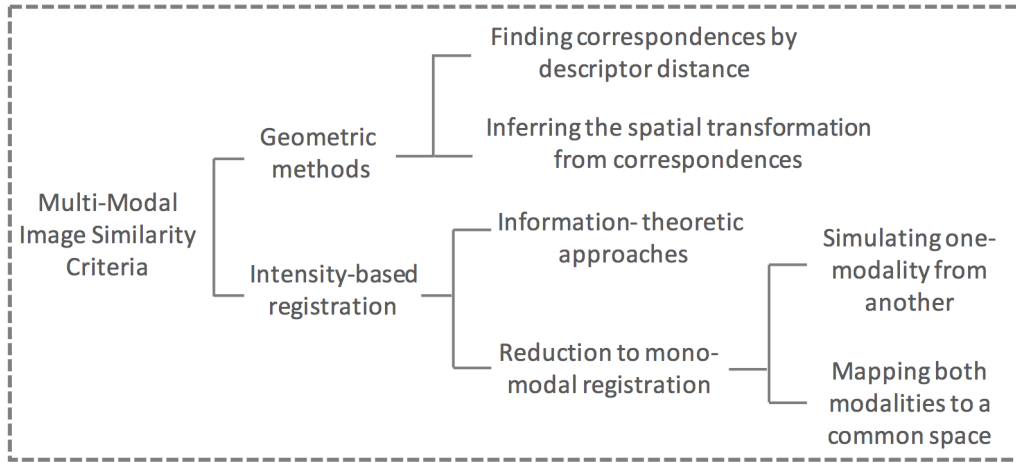


Figure 2.8: Categorisation of image similarity metrics that are used in multi-modal medical image registration problem.

different intensity statistics and the relationship between the two image signals might vary depending on the spatial location and scanner configuration (e.g. US probe location). These factors are more pronounced particularly in the cardiac US multi-modal image registration problem. The focus of this review and thesis will be on the development of reliable image similarity metrics to identify correspondences between multi-modal images (Figure 2.7).

The diagram shown in Figure 2.8 visualises the standard categorisation of the existing similarity criteria used in multi-modal registration methods. Information theoretic measures such as mutual information (MI) [167] and correlation ratio [221] have been common choices in alignment of multi-modal MR image data. Nevertheless, in US image analysis, global similarity metrics do not always capture the true tissue correspondences since image intensity statistics can vary spatially. For this reason, the focus has shifted to development of local intensity similarity metrics such as local MI [253] and conditional MI [151]. King et al. [127] proposed a physical model based US dense similarity metric and used to align CT and US cardiac images. Other methods, on the other hand, have approached the multi-modal registration problem by converting it to a mono-modal registration problem. This goal can be achieved either by simulating one modality from another or by mapping both modalities into a common space.

Simulation methods have been used to align abdominal CT/US [270], fetal MR/US [133], and cardiac US/CT/MR data [116]. The main drawback of the methods in this category is that

they require tissue segmentation prior to registration. Furthermore, the intensity models used in modality simulations may not reflect the true intensity characteristics of the scanner. For this reason, research has focused on techniques that map the images into a common feature space. Multi-channel feature based alignment techniques were the early examples, such as wavelets coefficients [276] and Gabor features [196]. A particular feature type, namely local phase representation, was used as a similarity metric in the cardiac MR/US registration problem [284]. The proposed approach extracts ventricle boundaries from both US and MR images by analysing the phase signal in a multi-scale fashion. Recent work [107] has shown improved performance over these approaches by proposing the use of local self-similarity measures. It has achieved state-of-the-art performance in CT lung and US/MR brain image registration [109]. Although these methods have achieved good performance in benchmarks, their translation to clinical use is still limited due to the low US imaging quality. In particular, the generated representations may not always be robust towards noise and the registration accuracy strongly depends on the image quality. To overcome this problem, research needs to focus on development of machine learning based image similarity techniques. A trained model can be utilised to map inter-modality data into a common feature space to highlight the tissue correspondences.

In addition to intensity-based metrics, sparse geometric descriptors have been used in multi-modal image alignment as a similarity metric. Typically two sets of landmarks or salient points are extracted and these are matched via descriptor distance or geometric constraints such as graph matching algorithms [145]. For instance, SIFT feature based 3D MR/CT image alignment [262] is one particular example. Similarly, surface based cardiac US/MR image alignment techniques [213, 272] have been proposed. Procrustes analysis [102] has been commonly used to align surface point sets, which require prior surface segmentation for image alignment.

2.3.3 Evaluation of Image Registration Algorithms

The evaluation of registration algorithms is another important subject of study. Since the underlying ground truth spatial transformation information is not always available, except for

phantom data, several techniques have been developed to quantify the accuracy of registration algorithms. Rohlfing [224] showed that surrogate measures (e.g. SSD or MI) may not always be a good indicator of spatially meaningful registration results. For this reason, quantification based on spatial distance of physically attached markers is a more reliable way to evaluate registration algorithms [81], which is referred as fiducial localisation error. Similarly, it can also be achieved through sparsely located landmark correspondences or through comparison of object surface segmentations performed on both target and source images [224]. However, manual identification of landmarks or fiducial points is a tedious and error-prone task even for clinical experts. Therefore, inter-observer variability in landmark localisation accuracy often has to be taken into account in registration algorithm evaluation.

2.4 Machine Learning Models in Image Analysis

An overview of two machine learning models commonly used in image analysis is given in this section. In particular, we discuss machine learning approaches based on decision forests and convolutional neural networks. Our analysis focuses on model parametrisation, design rules in image analysis frameworks, and practical requirements in terms of training data size and computation power. Moreover, we compare the two models in terms of how they make use of training data to learn decision rules and reason about it. Finally, we summarise their applications in cardiac image analysis and discuss about the on-going practical challenges: model interpretability and computation resource management.

2.4.1 Ensemble of Decision Trees (Decision Forests)

The decision tree model [30] is a non-parametric model commonly used for classification and regression tasks. The model predicts the value of a target variable by using simple decision rules based on comparing the input data features to a fixed threshold value. As shown in Figure 2.9, the tree model is composed of split and leaf nodes, which define the tree model parameters. Split

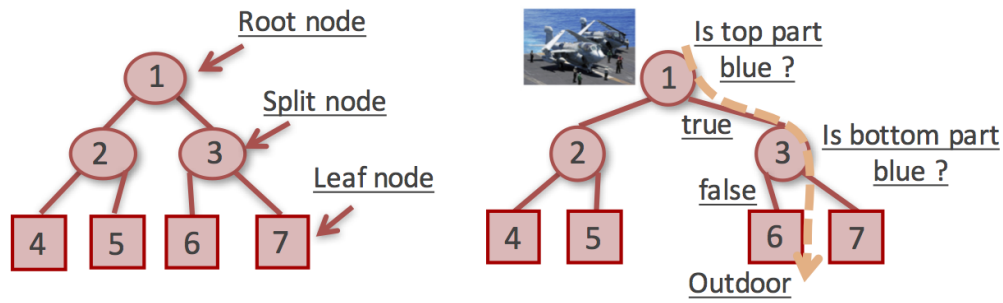


Figure 2.9: The definition of nodes in a decision tree structure. Image class label prediction is decided by routing the samples through the split nodes based on sequential decision rules.

nodes characterise the decision rules applied on input data, branches represent the outcome of each decision, and leaf nodes contain target posterior information for given input. As explained in the recent survey by Criminisi [57], decision tree training algorithms allow us to come up with an optimal set of sequential decision rules, used in split nodes, to minimise tree transversal and maximise information about output label for given input data, which is also known as minimisation of output space uncertainty or entropy.

The learned decision rules are stored in split nodes (weak classifiers), and in the standard tree model they are based on a decision stump. In a simple classification task, shown in Figure 2.10, the tree decision boundaries are composition of horizontal and vertical lines defined along a single dimension in feature space due to decision stumps in each split node. However, in canonical correlation based splits [214] and oblique splits [176], split node decision rules can be based on multi-dimensional information obtained from training samples, which can reduce model bias (underfitting) in training. At testing time, each sample is routed to its corresponding leaf node through learned decision rules as shown in Figure 2.9. Decision stump threshold values are determined by considering the information gain after each binary split of training data. The formulation of different information gain objectives are discussed in [58].

Decision trees offer design flexibility as such different kind of labels (e.g. numerical, categorical) can be utilised simultaneously within the same model as long as training data can be clustered in a meaningful way. Some examples include structured learning [67] and multi-task Hough Forest models [90]. Similarly, the local leaf node models can store different sorts of information

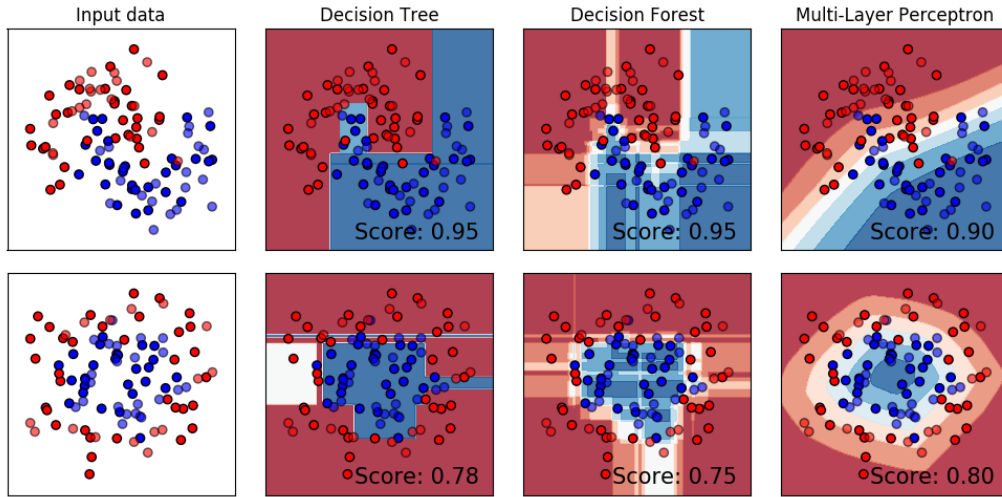


Figure 2.10: Visual comparison of different classification models and their decision boundaries. In this toy classification example, we observe that decision forests and MLP models generalise to the underlying data distribution in different ways.

(regression, classification, image meta-data) at the same time. In several studies [59, 258], the Bayesian uncertainty of a tree model is stored in leaf nodes to understand model prediction confidence. Another useful feature is tree pruning, through which model over-fitting can be avoided by simply reducing the depth of trained decision trees. For these given reasons, decision trees and ensemble models have been widely studied in image analysis in the last decade as it provides design flexibility.

Ensemble of decision trees [113], also known as decision forests, combine several uncorrelated decision trees into a single model through bootstrap aggregating (bagging) [29], where different subsets of training data and features are used to train each tree. In this way, the model variance is reduced and better data generalisation is achieved. In ensemble models, the final outcome is usually obtained by aggregating individual tree predictions, and weighted averaging is one way to merge multiple predictions as in the case of regression forest [59]. Another advantage of decision forests is that they are simple to understand and interpret. By analysing the information gain obtained at each split node for the selected image feature, we can categorise input features in terms of their importance for the given task. Additionally, by analysing the leaf nodes and adjacency of each sample stored in the leaf nodes, one could get insights about the local leaf models and visualise how samples are clustered within the tree model.

For example, a proximity analysis [31] is used in [190] to visualise how the cardiac images are clustered. Lastly, decision tree models do not require significant computation power and large amounts of data unlike convolutional neural networks which will be discussed next.

2.4.2 Convolutional Neural Networks

A convolutional neural network (CNN) [139] is a feed-forward artificial neural network model that has been successfully applied in image classification benchmarks such as ImageNet [132]. The neural connectivity pattern of a CNN model was inspired by the biological processes in the human visual system [117] and it was initially proposed by Fukushima and Miyake [88]. From a technical perspective, CNN models are parametrised by convolutional filters (learnt weights) that are shared across each neuron within the same layer, which reduces model complexity and exploits the statistical correlation between neighbouring pixels in an effective way. Therefore, the number of model parameters can be reduced significantly and the training of deep models become practically easier [100]. The connectivity pattern of CNNs is shown in Figure 2.11.

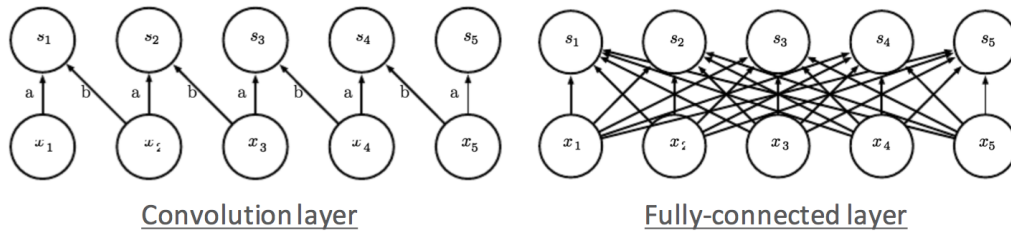


Figure 2.11: Neural connectivity comparison between convolutional and dense layers. In the former case, the connections are sparser and they are shared across the neurons, which reduces the number of parameters used in the model and partially avoids the over-fitting problem. The figure is adapted from [100].

A cascade of convolutional layers build a CNN model, as shown in Figure 2.12, which was initially used for digit recognition in the MNIST dataset [139]. The back-propagation algorithm [229] is used to learn the optimal set of image filters for the given task and training objective such as classification or regression. The updates on each layer parameters are computed by back-propagating the error signal originating from the loss function. The error signal is differentiated with respect to the weight and bias terms using the chain rule [100]. As long as the loss

function can be differentiated with respect to model parameters, the error originating from the loss function can be back-propagated to the input layer to optimise the model parameters. Due to non-convex nature of commonly used objective functions, CNNs are trained iteratively with stochastic optimisation methods. At each iteration, gradient updates are computed using a small subset of training data also known as mini-batch, which provides tolerance to noisy data and reduces the chance of the network getting stuck in local minima [271]. Additionally, iterative mini-batch update scheme enables the use of large scale datasets for model training.

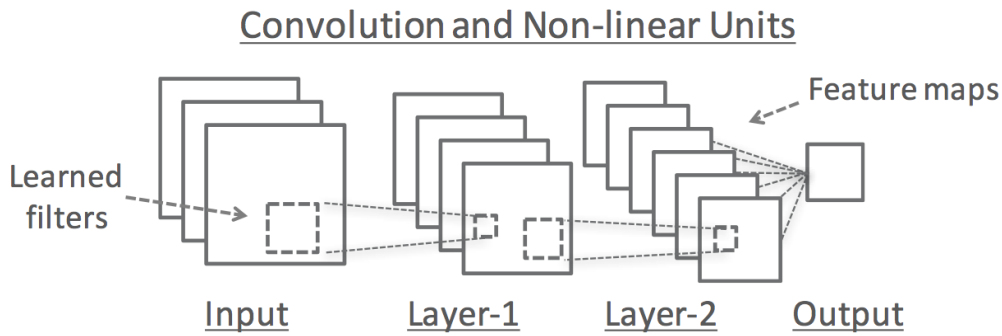


Figure 2.12: Structure of a standard convolutional neural network model that is composed of series of convolution layers and non-linear units. Each layer is characterised by a set of image filters that are trained to map the input data to the desired output space.

End-to-end learning and back-propagation also reduce the dependency of CNN models on hand-engineered image features and ad-hoc preprocessing techniques. This could be seen as the main advantage of CNN models over its counterparts as the generated intermediate representations are customised for the given training objective as shown in Figure 2.15. Indeed, this is the main reason why some studies [287, 288] proposed the use of auto-encoder models to extract domain specific image features from unlabelled image datasets. The learnt features are later used in an ensemble model (e.g. decision forest) for classification and regression tasks. Similar to DFs, in neural network models ensemble learning can be achieved using the drop-out technique [249] that is commonly used in densely connected layers to avoid model over-fitting. Due to its sparse connectivity structure, drop-out technique has not provided a substantial performance improvement when it is used in the CNN models.

Recent research has focused on development of CNN based image analysis frameworks, and

these approaches have consistently outperformed other ML models (e.g. decision forests) in common benchmarks in both computer vision [132] and cardiac imaging [237]. It is also worth to mention that from a theoretical perspective, Sethi [238] showed that any decision tree model can be approximated by a sparsely connected two-layer perceptron. Also, it is shown [246] that deeper network architectures are better at capturing non-linear mappings between input and output spaces for fixed amount of model capacity (number of trainable parameters). From these perspectives, one can argue that CNN models can potentially outperform tree models in image analysis tasks since CNNs can learn more complex mapping functions through a large number of hidden layers and non-linear operations. In the next section, we demonstrate how these models have been applied in medical image analysis tasks.

2.5 Applications of ML Models in Medical Imaging

Traditional image analysis techniques such as graph-cut and level-set segmentation methods operate according to a handcrafted optimisation objective that drive segmentation algorithm. Ventricle boundaries are delineated along high image contrast based on intensity data and some regularisation term that ensures smoothness and connectivity of predicted labels. However, these methods have underachieved in segmentation benchmarks (e.g. PASCAL VOC'12 [78]) in comparison to recent submissions done with neural network frameworks since no external learning and algorithm supervision is involved. In the medical imaging domain, one can observe a similar performance difference in the recent cardiac segmentation challenge [237]. With the advent of increased medical imaging data and annotations, the focus has shifted towards development of learning and pattern recognition based image analysis approaches that map input data into abstract representations to better exploit the connection between input and output spaces. As it can be seen in Figure 2.13, the interest in development and application of machine learning (ML) techniques has exponentially increased over the last decade.

Although most of the ML models, such as CNNs [139] and decision forests [113], were initially

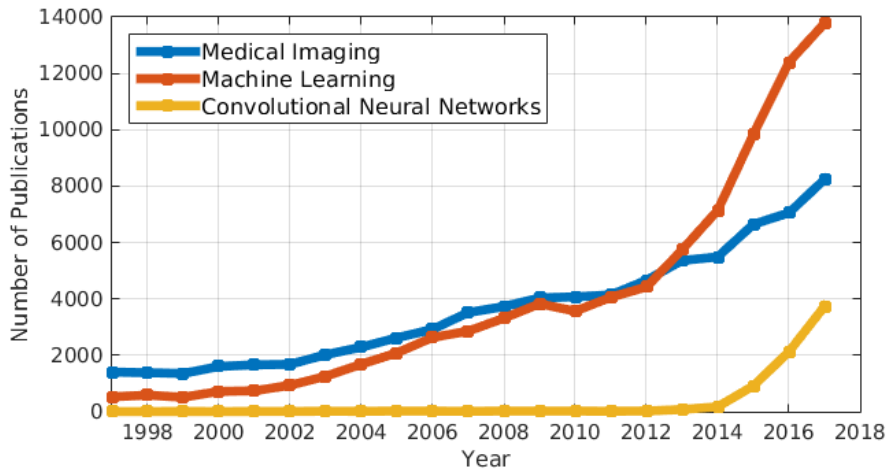


Figure 2.13: Number of scientific publications according to topic search on <https://apps.webofknowledge.com>. The figures show that attention to machine learning research has increased exponentially in the recent years. More importantly, convolutional neural network studies constitute substantial part of the machine learning research lately.

proposed long time ago, their widespread use in image processing frameworks has been delayed due to unavailability of large training data and computation power. Indeed, neural networks have seen a significant progress in the last few years and been commonly used in image processing pipelines since the availability of large computation power. The recent survey [104] presents a historical perspective on neural networks and their widespread use in medical image analysis.

In the context of cardiac imaging, researchers have trained various types of discriminative and generative models using imaging data: boosting trees, support vector machines, sparse coding and dictionary learning. The applications include semantic segmentation [173], landmark localisation [161], pathology classification [118], image reconstruction [234], and super-resolution [23]. The main factors that determine the performance of these ML approaches have been (I) the availability of training data, (II) the representation power of the model, and (III) the ability for generalisation of the trained model to different input data distributions such as images obtained from different scanners. The requirement of training data has been the main limitation of learning based approaches; however, in the last years more and more cardiac image datasets have become available such as UK Biobank [206] and Second Annual Data Science Bowl [237].

On such large scale datasets, the main performance difference between ML based image anal-

Table 2.1: Applications of decision forest in medical image analysis.

Reference	Model	Application; remarks
Criminisi et al. [59] [MICCAI'09]	Regression Forest	Bounding box localisation of organs in CT abdominal images
Lempitsky et al. [144] [FIMH'09]	Classification Forest	Cardiac ultrasound LV segmentation using standard decision forest.
Margeta et al. [173] [STACOM'11]	Classification Forest	Spatio-temporal box features are used to segment cardiac cine-MR images.
Alexander et al. [2] [MICCAI'14]	Regression Forest	Single image super-resolution method: applied to diffusion MR images.
Lombaert et al. [153] [MICCAI'14]	Laplacian Forest	Decision trees are used to select neighbouring atlases in multi-label propagation. It is an efficient way of clustering and selecting images for analysis.
Milletari et al. [177] [MICCAI'15]	Hough Forest	Cardiac ultrasound LV segmentation. Regression votes are used to identify relative location of each patch, which are later utilised to propagate and merge labels from multi-atlas dataset.
Tarroni et al. [260] [FIMH'17]	Structured Forest	A framework to control quality of cardiac cine-MR images. The trained model is used to verify the slice coverage of acquired 2D image stacks.

ysis algorithms arises from how they make use of the annotations to generalise and learn a discriminative or generative model. Additionally, smart preprocessing and augmentation techniques play an important role as well, which are discussed in detail in [150]. In that regard, convolutional neural network (CNN) models have dominated the medical image analysis field in the last years since they can learn complex and task-specific abstract representations to model the relationship between input and output. Moreover, they can scale to large training images well and do not require large number of model parameters in comparison to multi-layer densely connected neural networks. Therefore, the differentiating point between ML models arises from compact parametrisation of training data while having a strong representation power. Ensemble models have also shown good generalisation power on different medical imaging data; as such, decision forest (DF) based frameworks were proposed by Criminisi et al. [58] to analyse cardiac CT and MR data in various problems. In the next section, we discuss about the existing work on DF and CNN based cardiac imaging.

Table 2.2: Variants of decision forest model have been developed to solve different natural image analysis tasks. This includes joint classification and regression training, and structured learning of the output space.

Reference	Model	Application; remarks
Gall et al. [90] [TPAMI'11]	Hough Forest	Pedestrian detection and localisation in natural images. Joint training of classification and regression objectives.
Shotton et al. [245] [CVPR'11]	Regression Forest	Human body joint localisation in Kinect depth images for pose estimation.
Dollar and Zitnick [67] [ICCV'13]	Structured Forest	Object boundary map extraction in natural images. Structured forest stores segmentation patches in its leaf nodes instead of storing label posterior distributions.
Glocker et al. [97] [ECCV'12]	Classification Forest	Multi organ segmentation in abdominal CT scans. Regression and classification objectives are used together to provide model supervision.
Schulter et al. [235] [CVPR'15]	Regression Forest	Single image super-resolution approach. Leaf nodes store local linear prediction models to synthesise high resolution image patches.

2.5.1 Decision Forests in Medical Imaging

In Table 2.1, we provide a short list of work on cardiac image analysis that have utilised decision forests in their pipeline. It is important to note that the existing work that has applied DF in medical imaging, are not only limited with the ones described in the table. What we observe is that researchers have initially utilised the DF model in default applications such as object localisation and label classification tasks. Over the following years, the focus has shifted to applications in other domains such as super-resolution or quality-control analysis. Moreover, we also observe that design flexibility of DF model has enabled researchers to make modifications in DF model structure to improve performance. A few examples are: structured learning [192], joint classification and regression training objectives [97], hashing forest [50]. The same research pattern can be observed in natural image analysis studies as well which is shown in Table 2.2.

Besides the discussed properties of DF models, there are a few limiting factors for their wide use in medical imaging: (I) DF models do not always scale well to a large number training

data since batch of samples are required to be processed simultaneously during training of each split node; whereas, CNN models can circumvent this problem through stochastic learning with small subsets of training data using mini-batches. For example, this could be a limiting factor in training of a DF model with thousands of annotations from the UK Biobank dataset [206]. (II) DF models are usually required to be retrained from scratch when new training data is acquired. (III) The flow of information from input to output space is designed to be deterministic and each sample usually ends up in a single leaf node; however, in CNN models this flow can be branched into several paths through inter-leaved parallel neural connections. In that regard, the DF model representation power is limited. Besides these aforementioned aspects, in the next section we provide a more detailed comparison between CNN and DF models. Technical details about the structured and regression trees are provided in the following chapters.

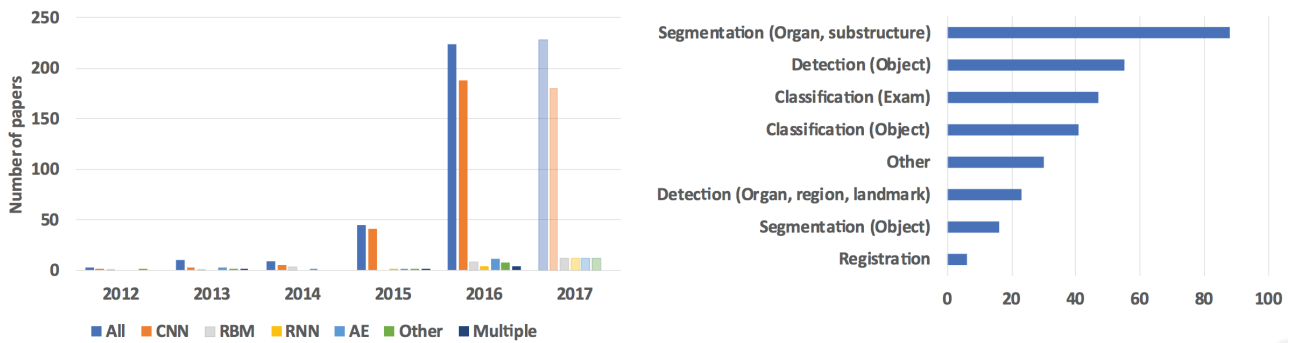


Figure 2.14: The bar-plot shows the number of medical image analysis publications over the recent years which are related to neural networks. On the right, the distribution of these publications with respect to the application type is shown. Segmentation and CNNs have been the most commonly studied research areas in the recent years. Image courtesy Litjens [150].

2.5.2 Convolutional Neural Networks in Medical Imaging

Research on convolutional neural networks [100] have shown that CNN models can learn complex decision boundaries through cascaded convolution and non-linear operations. Also as discussed in the previous sections, based on the recent benchmark results, we can expect that CNN models are more suitable for medical image analysis including cardiac MR images as long as sufficient training data is provided. Figure 2.14 shows the number of publications in medical

imaging that utilises a CNN model in the proposed analysis framework, which has been growing exponentially. More importantly, majority of the studies have focused on applying the CNN models for the image segmentation task, particularly the U-Net architecture [225] has been widely used in this problem.

As in the case of decision forest, the scope of the recent CNN based medical imaging research has shifted towards other applications such as: cardiac image quality assessment [282], MR image super-resolution [191] and image reconstruction [234]. In addition to these application areas, recurrent models and adversarial learning techniques [101] have attracted interest in medical imaging, and have been applied in CMR segmentation problem [210, 275]. Recurrent models (e.g. LSTM [114]) allow feed-forward and backward information passing between neurons by storing features from past model predictions in its hidden states, which combines both current input and past sequence information to determine model output. Adversarial learning [101], on the other hand, allows formulation of custom training objectives, which may not be easily formulated in closed-form. In this thesis, the last two chapters are dedicated to CNN based cardiac image analysis and the target applications are image super-resolution and segmentation. The related CNN work on these topics are further discussed below:

Medical Image Segmentation: A representative list of existing segmentation work that utilises a neural network model in its processing pipeline is given in Table 2.3. Early work [182] incorporated deep belief network models (cascaded Boltzmann machines) in feature extraction and nearest neighbour patch search problems. Similarly, Avendi et al. [6] utilised an auto-encoder model to obtain a coarse segmentation map of cardiac LV, which is later fine-tuned with a deformable model. Follow-up studies [45, 264] have performed end-to-end learning between input and label spaces using a fully-convolutional model (FCN) [154], which has outperformed the previous hand-engineered segmentation pipelines including multi-atlas and graph-cut techniques on the MICCAI'12 RV segmentation challenge data [207].

The medical imaging community has also explored the performance difference between 2D and

Table 2.3: A list of cardiac MR and ultrasound image segmentation frameworks that utilise deep neural network architectures.

Reference	Model & Modality	Application; remarks
Nascimento et al. [182] [ISBI'16]	DBN Model, Ultrasound	Deep belief network is trained on patches to guide multi-atlas ultrasound LV segmentation process
Chen et al. [45] [MICCAI'16]	CNN Model, Ultrasound	Multi-domain cardiac ultrasound segmentation model (2-,3-,4-CH views). Each branch of the model shares a common set of features and benefit from inter-domain knowledge.
Milletari et al. [178] [MICCAI'17]	CNN Model, Ultrasound	An externally trained statistical PCA shape model is used to guide CNN ultrasound LV segmentation.
Rupprecht et al. [230] [Arxiv'16]	Active Contour & CNN, MRI	CNN model is used to predict dense flow fields which guide active-contour fitting. The framework is used to segment LV in CMR images.
Avendi et al. [6] [MEDIA'16]	Stacked AE, MRI	Autoencoder is used to perform coarse LV segmentation which is later fine-tuned using a deformable model and level-set optimisation.
Poudel et al. [210] [MICCAI'16]	RNN Model, MRI	Recurrent model is utilised across through-plane direction and combined with in-plan CNN model, which is used for CMR LV segmentation.
Tran [264] [Arxiv'16]	CNN Model, MRI	Fully convolutional network is trained in an end-to-end manner for CMR LV segmentation

3D image kernels used in CNN based cardiac segmentation frameworks [15]. Since common benchmarks (e.g. ACDC'17¹) have been evaluated according to manual annotations which were done on each slice independently, 3D CNN models have not demonstrated substantial performance improvement. Besides the CNN architectures, Poudel et al. [210] has shown that non-feedforward models such as recurrent networks can be used to explore spatial consistency across slices without requiring 3D kernels.

Image Super-Resolution and Other Applications: Single image super-resolution (SR) is another application area of CNN models, which is discussed extensively throughout this thesis. Table 2.4 gives a short summary of the CNN based SR studies in natural and medical

¹<https://www.creatis.insa-lyon.fr/Challenge/acdc/>

Table 2.4: Single image super-resolution frameworks that make use of a trained CNN model. The application areas range from natural images to medical images including cardiac MRI [191] and brain diffusion imaging.

Reference	Model & Modality	Application; remarks
Dong et al. [71] [ECCV'14]	CNN Model, Natural images	Single image SR on natural 2D images. One of the earliest work that utilised a CNN model for single image SR problem.
Shi et al. [242] [CVPR'16]	CNN Model, Natural images	Single image SR on natural 2D images. Sub-pixel convolution technique is proposed to reduce model runtime by an order of magnitude.
Ledig et al. [141] [CVPR'17]	CNN Model, Natural images	Single image SR on natural 2D images. Adversarial learning is used in SR training to synthesise realistic looking high resolution images at test time.
Tanno et al. [259] [MICCAI'17]	CNN Model, Diffusion MRI	Single image super-resolution for brain DT images. Empirical Bayesian SR uncertainty is estimated using the drop-out technique.

image analysis. The early work of Dong et al. [71] has proposed a SR solution that uses a 3-layer shallow network. The approach was evaluated on natural images, and it has outperformed previous state-of-the-art SR techniques by a significant margin in terms of both PSNR and SSIM [269] metrics. Later, Shi et al. [242] has proposed a sub-pixel convolution method as an improvement to Dong's architecture. The new SR model performs all convolution and non-linear operations on the low-resolution image grid to reduce the computational complexity of the network.

As an application of similar models in medical imaging, a multi-input SR framework is proposed in [191] which utilises both LAX and SAX stack cardiac data to synthesise a high-resolution isotropic image volume. In follow-up work, Tanno et al. [259] have recently shown that with the same CNN model, SR regression uncertainty can be obtained empirically by sampling from the posterior distribution, which is often done with multiple forward-passes and drop-out at each forward-pass as suggested in [89]. Lastly, Ledig et al. [141] applied adversarial learning to SR CNN models to obtain qualitatively more appealing HR natural image synthesis. However, the experimental results show that the improved image quality may degrade the performance

Table 2.5: A list of other application areas of CNN models in cardiac imaging, which includes quality control, landmark localisation, and deformable registration.

Reference	Model & Modality	Application; remarks
Le et al. [138] [DLMI'17]	CNN Model, Cardiac MRI	Automatic anatomical landmark localisation for cardiac long-axis view generation.
Zhang et al. [282] [SASHIMI'16]	CNN Model, Cardiac MRI	Automatic image quality assessment. The trained model identifies whether the apical or basal slice is missing in the acquired stack data.
Zhang et al. [158] [MICCAI'17]	CNN Model, Ultrasound	FFD based myocardial tracking method is regularised with a spatio-temporal CNN model
Ghesu et al. [92] [MICCAI'16]	Reinforcement Learning, Cardiac MRI	Artificial agent based automatic anatomical landmark localisation, which is trained using Q-Learning.
Rohe et al. [223] [MICCAI'17]	CNN Model, Cardiac MRI	Stationary velocity fields between two images is computed using a trained model to perform deformable registration.

in terms of PSNR.

Besides the segmentation and super-resolution approaches, CNN models have found other application areas in cardiac imaging. Deformable image registration using ML techniques is one of them, and it still remains a particular research area which has not been investigated in-depth. As shown in Table 2.5, recent work [158, 223] has utilised the U-Net model to approximate parameters of FFD and SVF deformation models. These techniques can implicitly learn the underlying pairwise image similarities and identify the tissue correspondences. In Chapter 3, we demonstrate that the proposed probabilistic surface representation is one way to explore multi-modal image correspondences through trained decision forests. However, CNN based approaches could potentially explore a more discriminative intermediate space where similarities between multi-modal data can be better identified.

Similarly, decision forest based cardiac landmark localisation techniques [190] have been benchmarked in [138] against a novel CNN architecture and performance improvements were reported. Similarly, for the same localisation problem, Ghesu et al. [92] has shown that an artificial agent trained with a reinforcement learning technique can locate the right ventricle insertion points

iteratively with even better accuracy. Another interesting application area is the automatic image quality control for cardiac cine MR image stacks. Zhang et al. [282] approached this problem using a CNN model and automatically verified slice coverage of acquired cardiac MR stacks with the trained models.

Incorporating Anatomical Priors in Neural Networks: As discussed in Section 2.1.4, anatomical priors have been incorporated in traditional cardiac segmentation algorithms to provide useful guidance in optimisation. However, the integration of shape and label prior information into CNN models have not been studied in depth in medical imaging except for few studies: boundary-aware segmentation [19], topology-aware CNN segmentation [44]. In recent work, principal component analysis [178] regularisation model has been incorporated in CNN models to provide shape prior information in cardiac ultrasound segmentation problem. This subject is investigated in the last chapter with a more comprehensive survey on the related work published in computer vision field.

Current Limitations: Although CNN models can learn complex non-linear mappings and perform well in image analysis tasks, they are often described as black boxes. Especially in medical image analysis, accountability is important and it is often not enough to have an accurate prediction. Zeiler and Fergus [279] has shown that by analysing the intermediate feature-maps, we could have a better understanding of what each layer in the network is specialised at. Similarly, guided gradient back-propagation technique [248] was proposed as another way to visualise the salient image features that contribute to model prediction. However, future work is required to investigate the interpretability problem in more depth to understand what these models are actually learning. Besides the model interpretability, hardware limitations (e.g. GPU memory) are still another limiting factor in analysis of large image contexts particularly in 3D isotropic images (e.g. abdominal CT data). Recent studies have tried to address this problem by exploring large context sparse convolution techniques which reduce the computational demand and number of feature maps [108].

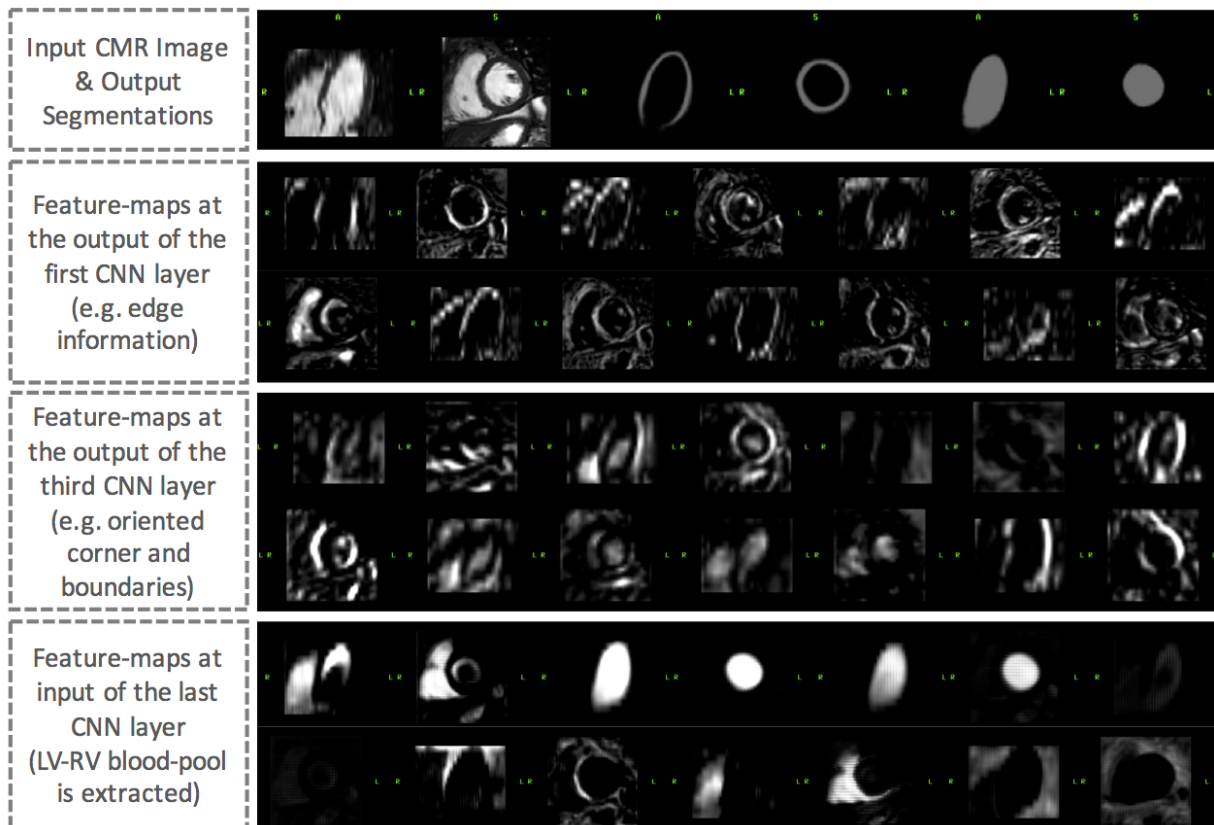


Figure 2.15: Visualisation of intermediate feature maps generated with a cardiac LV segmentation CNN model. The input image shown on the top figure is passed through a series of convolutional layers. In the early stages, the model extracts low level features such as the ventricle boundaries. In the following layers, oriented boundary information is computed. Towards the end, the model extracts semantic labels by aggregating the inferred contextual information.

Chapter 3

Multi-Modal Cardiac Image Analysis through Probabilistic Edge Maps

This chapter is based on:

- O. Oktay, A. Schuh, M. Rajchl, K. Keraudren, A. Gomez, M. P. Heinrich, G. Penney, and D. Rueckert, “Structured Decision Forests For Multi-modal Ultrasound Image Registration”, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2015.
- O. Oktay, A. Gomez, K. Keraudren, A. Schuh, W. Bai, W. Shi, G. Penney, and D. Rueckert, “Probabilistic Edge Map (PEM) for 3D Ultrasound Image Registration and Multi-atlas Left Ventricle Segmentation”, International Conference on Functional Imaging and Modelling of the Heart (FIMH), Springer, 2015.
- O. Oktay, G. Tarroni, W. Bai, A. de Marvao, D. O’Regan, S. Cook and D. Rueckert “Respiratory motion correction for 2D cine cardiac MR images using probabilistic edge maps” Computing in Cardiology Conference (CinC), IEEE, 2016.

3.1 Introduction

In this chapter, we present a novel image feature representation, namely Probabilistic Edge Map (PEM), and its application to cardiac image analysis tasks such as multi-modal image registration and segmentation. Image representations have been commonly used in image analysis and understanding to transform the raw intensity data to a new space, where values are intended to be more informative, robust to image noise, non-redundant, and in some cases leading to easier interpretation of images. A few well known categories are: low-level features (e.g. gradient magnitude, local phase [131], SIFT [157]), shape features (e.g. Hough transform), and motion features (e.g. optical flow [165]).

Similarly, the proposed PEM representation identifies the structure of tissues and organs in medical images by extracting their probabilistic surface representation as shown in Figure 3.1. In particular, in ultrasound (US) image analysis object boundaries are the main source of information since raw image intensities do not provide high soft-tissue contrast as in alternative modalities such as MRI. Furthermore, PEM based surface representations can guide algorithms to identify multi-modal correspondences since it is modality-independent, which can be used to perform spatial alignment of multi-modal images.

PEMs differentiate from classical image representation techniques by integrating a machine learning model in the feature extraction process. In that regard, PEMs are computed by passing input images through a structured decision forest [67]. By making use of training data, we are able to increase the algorithm's robustness against image noise, and also avoid spurious edge responses. Indeed, it is the main reason why PEMs provide well defined response compared to the other feature representation techniques considered in this study (Figure 3.1).

We demonstrate three different applications of PEM image representation for multi-modal cardiac image analysis: (I) multi-modal cardiac image registration, (II) multi-atlas segmentation of cardiac ultrasound images, and (III) respiratory motion correction of cardiac cine-MR images. All three applications make use of the PEMs by utilising them in image registration as

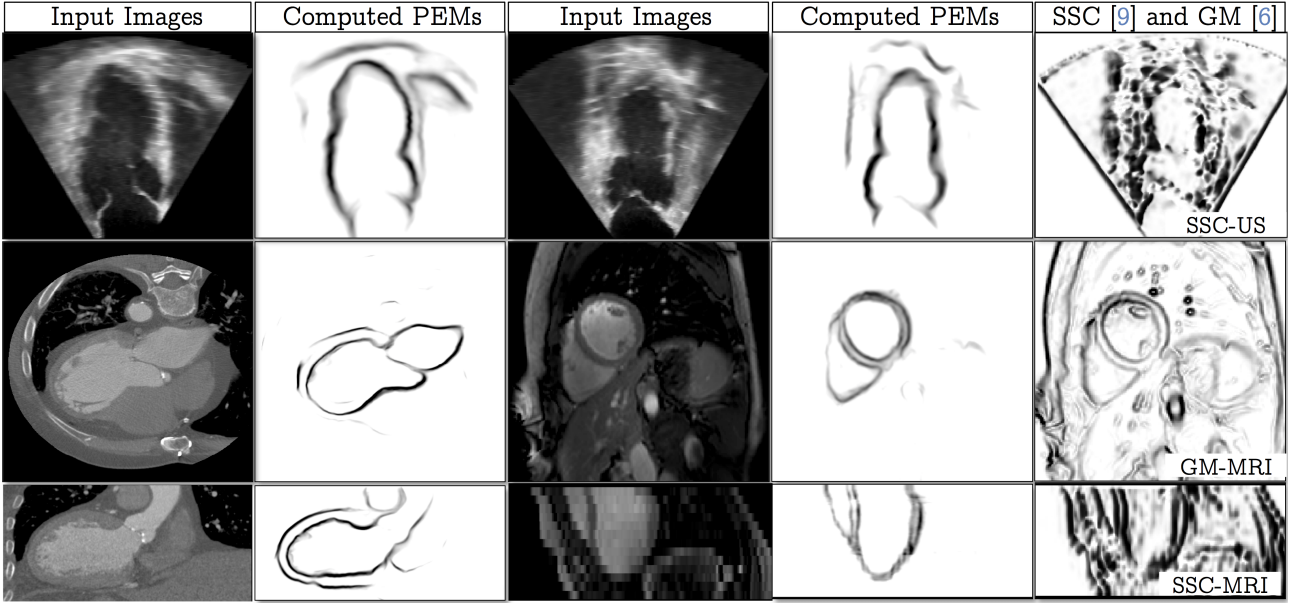


Figure 3.1: Example of PEMs, obtained from cardiac images of different modalities, are compared to gradient magnitude (GM) and self-similarity descriptors (SSC). Since SSC is a vectorial descriptor, only a single component of SSC is visualized.

a image similarity metric. At the end of the chapter, we also demonstrate the applicability of PEMs for other organs of interest such as abdominal images.

3.2 Probabilistic Edge Maps (PEMs)

PEMs capture probabilistic representation of organ surface and they can be configured to be target-organ specific. Furthermore, PEMs are modality independent; as such the same representation can be generated from different modalities (e.g. US, CT, or MRI), which requires training of a separate decision forest for each modality. However, a single training configuration can be applied to train the forest models for all modalities.

PEM representation is used as an intermediate space to identify correspondences between images from different modalities. In other words, multi-modal image similarity metrics are defined in the PEM space. In this way, subsequent multi-modal image analysis can be facilitated more easily. For instance, image registration optimisation can be performed solely based on the surface points of target organs without the influence of the background tissues. The proposed

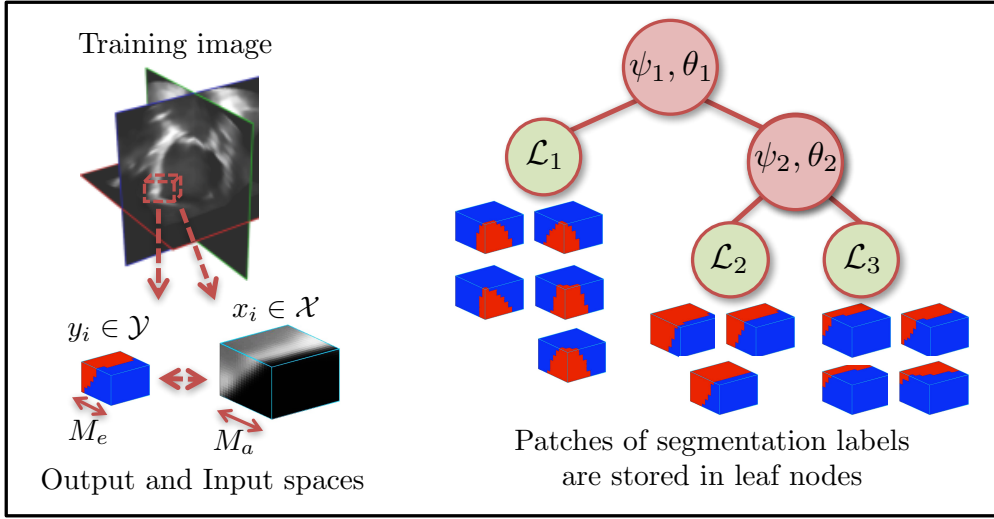


Figure 3.2: Structured decision forest: Tree leaf nodes store patches of segmentation labels instead of single voxel posterior distribution over class labels. The tree output space is structured and node splits are performed by mapping structured labels to one dimensional space, where standard information gain objective is maximised to determine split parameters.

representation type is motivated by the fact that B-mode US images focus mainly on the anatomical structure of the ventricle boundaries. In order to make use of this information in an effective way, the ventricle boundaries should be extracted from given imaging modalities. In contrast to the other structural representation types (Figure 3.1), PEMs utilise training label annotations and a trained boundary classification model. For this reason, it provides smoother and cleaner object boundary maps. From another perspective, PEMs can be considered as fuzzy boundary segmentation maps where subsequent analysis is performed.

This structural representation, PEM, is generated from a structured decision forest (SDF). While SDFs are very similar to decision forests, they possess several unique properties. In particular, in SDFs the output space $\mathcal{Y}$ is defined by structured segmentation labels $y_i \in \mathcal{Y}$ rather than a single label $y_i \in \{0, 1\}$. SDF decision trees can be trained as long as the high-dimensional structural labels can be clustered into two or more subgroups at each tree node split ψ_j as illustrated in Figure 3.2. This is achieved by mapping each patch label to an intermediate space $(\Theta : \mathcal{Y} \rightarrow \mathcal{Z})$ where label clusters can be generated based on their pairwise Euclidean distances in $\mathcal{Z}$ (cf. [67]). Once the training is completed, each leaf node $\mathcal{L}_j$ stores a segmentation patch label $y_i \in \mathbb{Z}^{(M_e)^3}$ of size $(M_e)^3$. Additionally, the corresponding edge map

$y'_i \in \{0, 1\}^{(M_e)^3}$ is stored as well. This allows predictions to be combined simply by averaging during inference.

Similar to decision forests, the input space $\mathcal{X}$ for the SDFs is characterized by the high dimensional appearance features ($x_i \in \mathcal{X}$) extracted from image patches of fixed size $(M_a)^3$. These features are computed in a multi-scale fashion and correspond to intensity values, gradient magnitudes, six HoG-like channels, and local phase features. The weak learner parameters (θ_j), e.g. stump threshold value and selected feature channel id, are optimized to maximize Shannon entropy $\mathcal{H}_1$ based information gain at each node split:

$$\mathcal{I}(S_j) = \mathcal{H}_1(S_j) - \sum_{k \in \{L, R\}} \frac{|S_j^k|}{|S_j|} \mathcal{H}_1(S_j^k) \quad (3.1)$$

Here $S_j = \{\mathcal{T}_i = (x_i, y_i)\}$ is a set of patch labels and features reaching node j . During the PEM generation (testing time), each input voxel accumulates $N_t \times (M_e)^3$ overlapping binary edge votes y'_i from N_t number of trained decision trees. These predictions are later averaged to yield a probabilistic edge response. PEM predictions are bounded in the interval of $[0, 1]$ and they store a soft-value for each voxel based on the edge strength.

Spatial aggregation of a large number of edge votes results in a smoother and accurate delineation of the structures of interest. It is similar to aggregation of class predictions obtained from multiple trees in a standard classification forest. However, PEM responses are not quantised to obtain a binary edge mask. Fuzzy object boundaries are indeed experimentally shown to be more useful for subsequent image analysis such as image registration optimisation. Additionally, PEM generation is performed on multiple image scales as suggested in [67]. The responses that are obtained at each image scale are aggregated using linear interpolation.

The segmentation label patch size $(M_e)^3$ has a direct influence on the boundary prediction accuracy; as such, smaller label patches yield increased false-positive predictions and spurious edge responses that can be observed in the background. With larger label patches, on the other hand, noisy responses are suppressed as the spatial structure of the output space is exploited

and PEM predictions become smoother. However, the complexity of the SDF model increases as the leaf nodes store large patches. Moreover, the clustering of samples at split nodes might tend to be less accurate due to a fixed amount of pairwise label comparisons. Image voxel resolution is another important aspect that has an influence on the accuracy of the generated edge maps. For instance, coarse through-plane resolution in CMR images can introduce bias in label split nodes. In order to resolve this issue, intensity images can be either resampled to near isotropic voxel resolution, or the label patch size can be adjusted by taking into account the voxel size differences. The next section describes how PEMs can be used in multi-modal medical image registration problem.

3.3 MR-US Cardiac Image Registration with PEMs

Interventional procedures in cardiovascular diseases often require ultrasound (US) image guidance. These US images must be combined with pre-operatively acquired tomographic images to provide a roadmap for the intervention. Spatial alignment of pre-operative images with intra-operative US images can provide valuable clinical information. Existing multi-modal US registration techniques often do not achieve reliable registration due to low US image quality. To address this problem, we propose the use of a novel medical image representation, named probabilistic edge map (PEM), as a similarity metric in image registration. PEMs generate similar anatomical representations from different imaging modalities and can thus guide a multi-modal image registration algorithm more robustly and accurately. The presented image registration framework is evaluated on a clinical dataset consisting of 10 pairs of 3D US-CT and 7 pairs of 3D US-MR cardiac images. The experiments show that a registration based on PEMs is able to estimate more reliable and accurate inter-modality correspondences compared to other state-of-the-art US registration methods.

Clinical motivation: In cardiovascular minimally invasive procedures such as mitral valve repair and aortic valve implantation [184], pre-operative surgical plans and roadmaps may be

complemented by intra-operative image guidance provided by ultrasound (US) and fluoroscopy. Computed tomography (CT) and magnetic resonance imaging (MRI) are widely used for planning of cardiovascular interventions since they can provide detailed images of the anatomy which are not usually visible in intra-operative cardiac US images. As the shape and pose of the heart changes over time during the intervention (e.g. due to respiration and patient motion), registration and fusion of pre- and intra-operative images can be a useful technology to improve the quality of image guidance during the intervention and subsequent clinical outcome. However, the difference in appearance of the heart in the different image modalities makes multimodal US image registration a challenging problem. In particular, the choice of an appropriate multimodal image similarity metric can be considered as the main challenge.

Related work: Current approaches for multi-modal US registration can be classified into two categories: (i) those using global and local information theoretic measures, and (ii) those reducing the multi-modal problem to a mono-modal problem. Early approaches in the first category suggested the use of normalized mutual information (NMI) [254] as a multimodal similarity metric. Recently, local similarity measures were proposed as an alternative to the global similarity measures. This reflects the fact that the US image characteristics show local variations due to wave reflections at tissue boundaries. Measures such as local NMI (LNMI) [128] and local correlation coefficient (LCC) [87] are designed to deal with this non-stationary behaviour. The second category of approaches converts the images from different modalities into a common feature space so that a mono-modal registration can be performed. In [133], fetal US and MR brain images are registered by generating pseudo-US images from MRI tissue segmentations. Other approaches, such as local phase [283], gradient orientations [63], and self-similarity descriptors (SSC) [109], generate similar structural representations from images to find correspondences between different imaging modalities.

Contributions: In this section we propose the use of probabilistic edge map (PEM) image representation for multi-modal US image registration. In contrast to local phase, SSC and

gradient magnitude [87] (GM) representations, PEMs highlight only the image structures that are relevant for the image registration as can be seen in Figure 3.1. This is because some of the anatomical structures visible in US images may not be visible in CT/MR images and vice versa, e.g. endocardial trabeculae, clutter, and shadowing artefacts. Therefore, extracting and registering only the left ventricle (LV) and atrial boundaries is a more robust approach than registering every voxel in the images. Moreover, PEMs generate a clear and accurate tissue boundary delineations, Figure 3.1, in comparison to other image representations. This improvement produces better spatial alignment and increases the robustness to noise which is important in US image registration. Additionally, PEMs are less sensitive under rigid transformations, unlike SSC descriptors. Thus they can cope with large rotational differences between images. Lastly, the proposed approach does not require tissue segmentations and can be generalized to other organs and modalities differing from the simulation based registration methods.

The proposed generic and modality independent representation is generated by a structured decision forest [67]. PEMs from different modalities are aligned with a block matching algorithm based on robust statistics followed by a deformable registration. The proposed registration framework is evaluated on US/CT and US/MR image pairs acquired from different pathological groups. The results show that PEM achieves lower registration errors compared to other image similarity metrics. This observation shows that PEM is a well-suited image representation for multimodal US image registration.

3.3.1 Methodology

Probabilistic edge map (PEM) representation: Cardiac US images mainly show the anatomy of the cardiac chambers and less detail of soft tissues. Therefore, further processing of these images, e.g. registration, requires a structural representation of the cardiac chamber boundaries such as the left ventricle (LV). In this work, a structural representation, PEM, is generated from a structured decision forest (SDF). Learning the image representation from training data increases the robustness in US images and allows the registration to focus on the

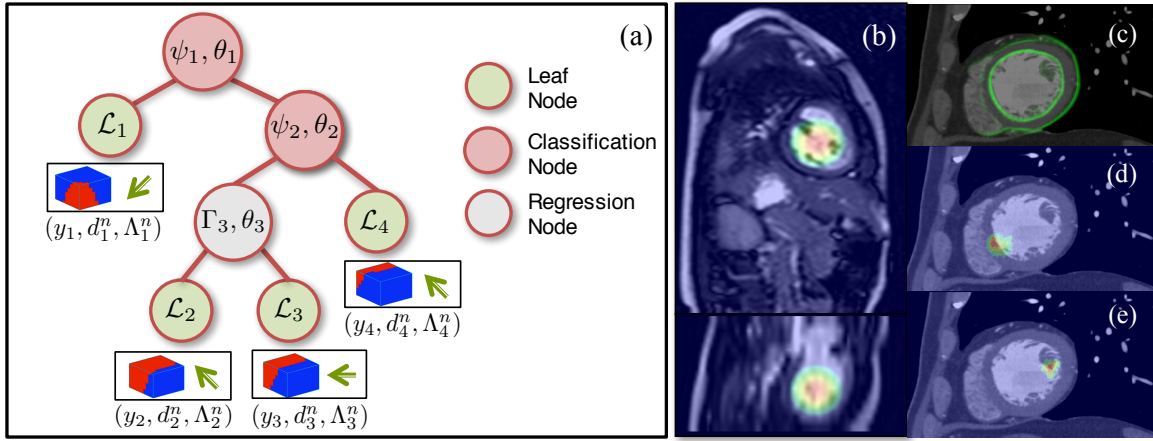


Figure 3.3: Structured regression tree: Leaf nodes contain segmentation patches and also spatial displacement vectors to anatomical landmark locations (a). Mid-ventricle (b), mid-septal (d) and mid-lateral (e) wall landmark localization by using PEMs (in green)(c) and regression nodes.

organs of interest. As can be seen in Figure 3.1, the PEMs outline only the LV and atrium while ignoring irrelevant anatomical boundaries. In addition, the generated boundaries are cleaner and smoother.

Structured regression forest: The presented SDF can be modified to allow additionally for simultaneous voting for predefined landmark locations such as the apex, mid-lateral and mid-septal walls, as shown in Figure 3.3. Each landmark point is detected independently in both target and source images. This can be used to obtain an initial rigid alignment as a starting point for the registration. Similar to Hough forests [90], classification nodes ψ_j are combined with regression nodes Λ_j in the tree structure. In this way, in addition to the patch labels y_i , each leaf node contains an average location offset d_i^n and confidence weight σ_i^n for each landmark $n \in \{1, \dots, N\}$ (cf. [56]). Inclusion of these nodes does not introduce any significant computational cost at testing time but adds additional information about the location of the landmarks.

The set of offsets $\mathcal{D}_i = (d_i^1, \dots, d_i^N)$ in the training samples $\mathcal{T}_i = (x_i, y_i, \mathcal{D}_i)$, is defined only for the voxels close to the boundaries of the ventricle. At each regression node Λ_j , the offset distributions are modelled with a multivariate Gaussian. The differential entropy of these

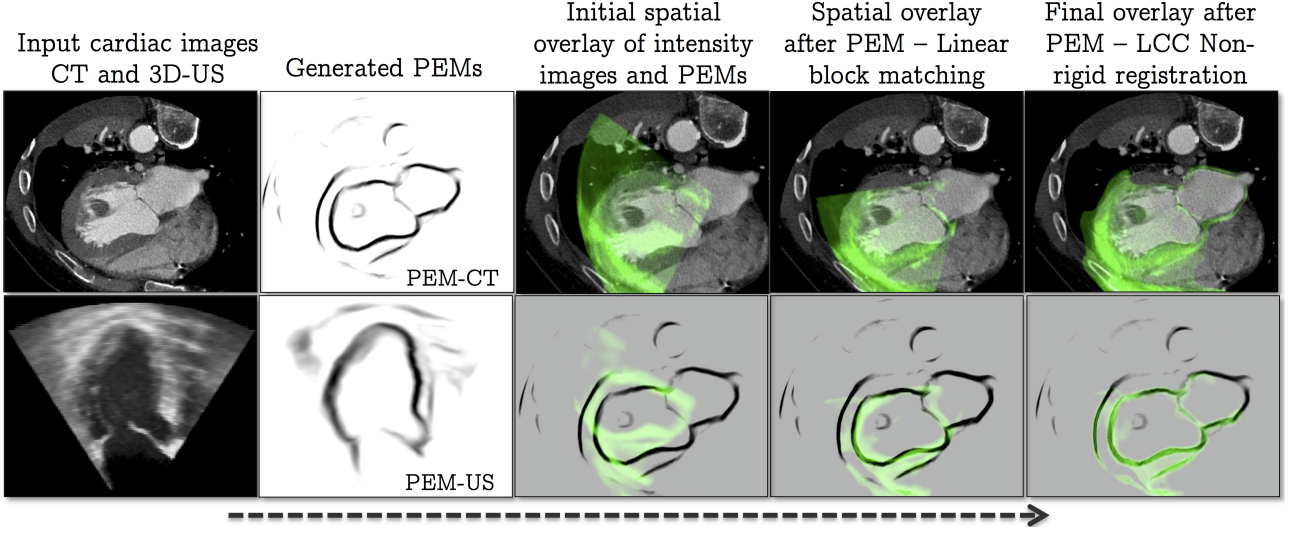


Figure 3.4: A block diagram of the proposed multi-modal image registration method.

distributions [56] is used as the uncertainty measure $\mathcal{H}_2$ in the training of Λ_j . In training, classification and regression nodes are selected randomly [90]. During testing, landmark location votes are weighted by the confidence weights and edge probabilities of the voxels.

Global alignment of PEMs: Similar to [133], a block matching approach is used to establish spatial correspondences between the PEMs (Figure 3.4), where the normalized cross correlation (NCC) is used as a measure of similarity. As suggested in [198], the set of displacement vectors computed between the corresponding PEM blocks are regularized with least-trimmed squared regression before estimating the global rigid transformation R . In this way, the influence of image blocks with no correspondences is removed, reducing the potential registration error introduced by shadowing or limited field-of-view in US images.

Non-rigid alignment of PEMs: To correct for the residual misalignment, due to cardiac and respiratory motion, between target (P^T) and source (P^S) PEMs, B-spline FFD [227] based non-rigid alignment follows the global rigid registration. The total energy function, minimized with conjugate-gradient descent optimization, is defined as:

$$E(T) = -LCC(P^T, P^S \circ T \circ R) + \lambda BE(T) \quad (3.2)$$

where λ is the trade-off between the local correlation coefficient [36] (LCC) similarity metric and bending energy regularization (BE). As the SDF classifier makes use of intensity values, local intensity variations in US images influence the edge probabilities in PEMs. For this reason, a local similarity measure is more suitable for PEMs than a global measure such as sum of squared differences. For similar reasons, LCC was used in [87] to align US-MR gradient magnitude images.

3.3.2 Experiments and Results

The proposed PEM registration framework is evaluated on 3D US-CT (10 pairs) and 3D US-MR (7 pairs) cardiac images. The CT images were acquired pre-operatively from adults after a contrast injection. The corresponding transesophageal US images were acquired during the cardiac procedure. In the second dataset, lower quality trans-thoracic US acquisitions are registered with multi-slice, cine-MR images (slice thickness 8 mm). Five of these pairs were acquired from children diagnosed with hypoplastic left heart syndrome and the rest from healthy adults. As a preprocessing step, all the images are denoised using non-local means [34] and resampled to isotropic voxel size of 0.80 mm per dimension.

In both experiments, PEM-LCC registration performance is compared with LNMI [128] and SSC [109] similarity based registration methods. Implementations of these methods were obtained from their corresponding authors. Particularly, in the SSC method, linear registration of the descriptors is performed by using least-trimmed squares of point-to-point correspondences for improved robustness. In the local alignment of SSC, a discrete optimization scheme (deeds [106]) is used to optimize SSC energy function. In LNMI, histograms with 64 bins are built locally and the registration is optimized using stochastic gradient descent. The number of bins is optimized for best performance.

In all three modalities, PEMs are generated and evaluated with the same training, testing and registration parameters. A PEM-SDF classifier was trained for each modality with set of cardiac

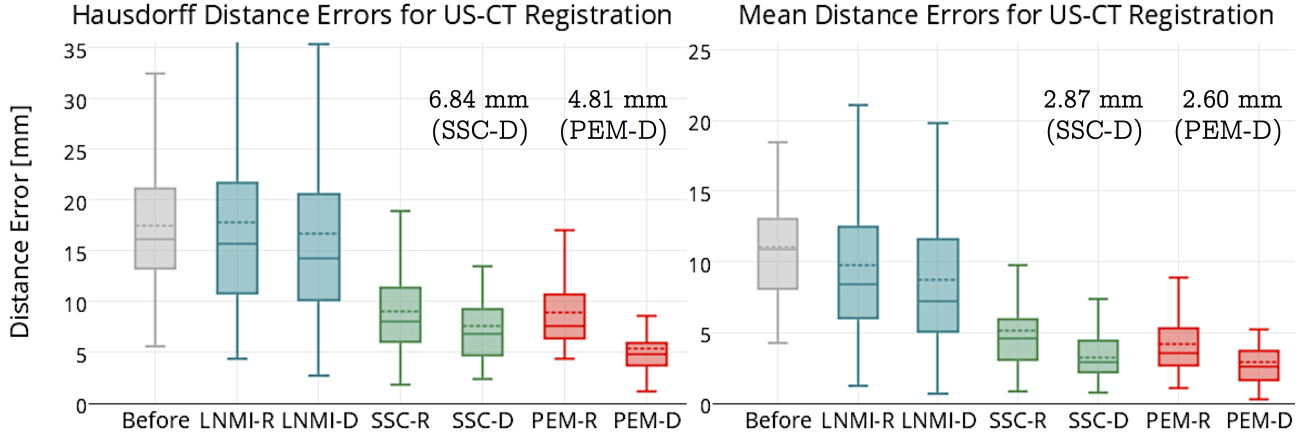


Figure 3.5: US/CT registration errors after rigid (R) and deformable (D) alignment. The mean and median values are shown in dashed and solid lines respectively.

images (50-80 images/modality) that is disjoint from the test set (17 pairs). It is important to highlight that the images from different modalities are not required to be spatially registered or come from the same subjects because the classifiers are trained separately for each modality. This significantly increases the availability of training data. The training data contains labels for myocardium, LV endocardium and atrium. The registration results are evaluated based on mean and Hausdorff distances of seven landmark points in the US images to closest points on the manually annotated LV endocardial surface in CT and MR images. The landmark points correspond to: apex, apical (2), basal (2), and mid-ventricle (2). For each method, the distances are computed after rigid and deformable alignment, and they are reported together with their values before the registration. For each image pair, 10 different registrations were performed with random global transformations for initialization.

US/CT evaluation results: The registration errors, provided in Figure 3.5, show that local intensity statistics based similarity measures do not perform consistently and in some cases the registration fails to converge to the correct solution. This suggests that structural representation methods, such as PEMs and SSC, are more reliable measures in US/CT image registration. On the other hand, after local alignment, PEMs achieve lower registration errors compared to SSC, which can be linked to the accurate and smooth boundary representation provided by PEMs. As shown in Figure 3.1, PEMs are less sensitive to noise and follow the anatomical boundaries

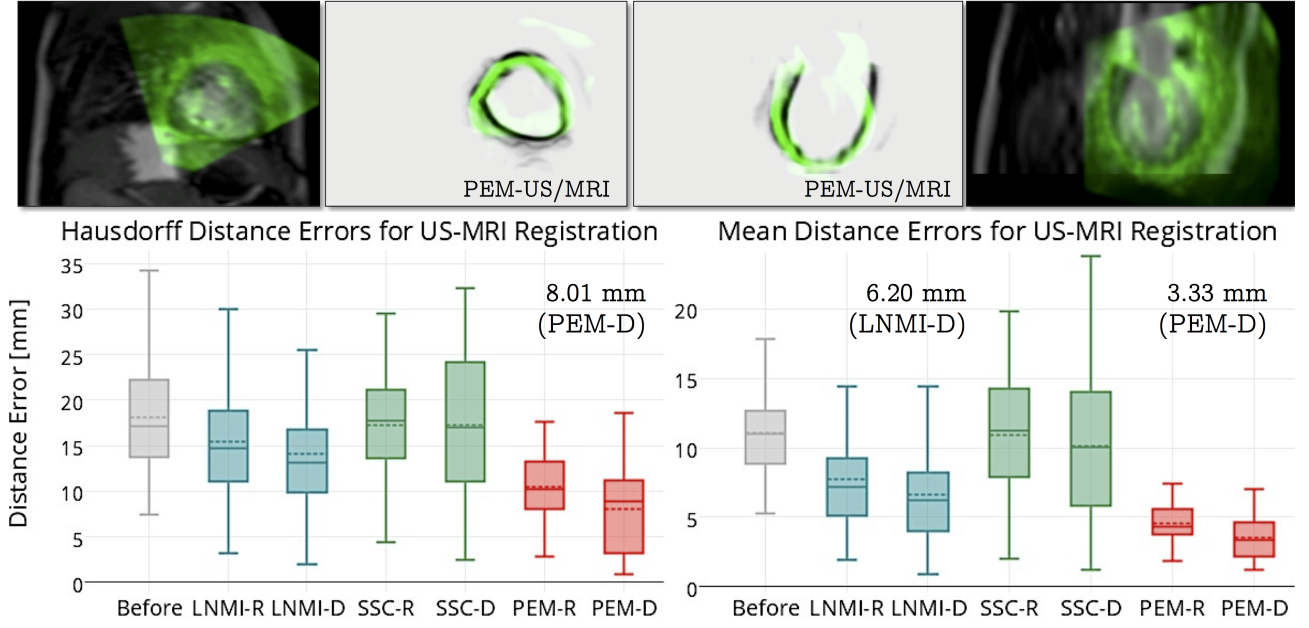


Figure 3.6: Global alignment of US/MR images (top). The boxplot shows the US/MR landmark distance errors after rigid (R) and deformable (D) alignment.

more accurately.

US/MRI evaluation results: It is observed that the SSC descriptors do not provide enough guidance in alignment of MR images acquired from children. As shown in Figure 3.6, the images contain more texture due to smaller size of the heart and trabeculae in transthoracic acquisitions. In contrast to this, PEM relies on annotations of the boundaries used in the training. It is able to generate a more accurate structural representation that is better matching with the myocardial boundary in MR images. The performance difference, shown in Figure 3.6, is explained with this observation.

Implementation details: All experiments were carried out on a 3.00 GHz quad-core CPU. The average computation time per registration was 73s for non-rigid registration, 21s for rigid alignment and 20s to compute each PEM. The training of the SDFs (70 min/tree) was performed offline prior to the registrations. It is important to note that PEM computation time can be reduced by parallelising the model inference over multiple trees and image patches.

3.3.3 Discussion and Conclusion

In the experiments after spatially aligning US/CT and US/MR images, it is observed that the endocardial annotations in US images underestimate the LV volume compared to the annotated CT and MR images. Similarly, in a clinical study [241], US volumetric measurement errors and underestimation were reported, and this situation was linked to the visible endocardial trabeculae in US images. For this reason, in our framework the US annotations are morphologically dilated before the training of the SDF. In this way, PEMs generated from US images become more consistent with the edge maps predicted from CT and MR modalities. In contrast, the other structural representation techniques are expected to fail to achieve optimal local alignment accuracy as these tissues are not visible in CT and MR images (Figure 3.6).

In this section, a novel PEM image representation technique and its application on US multi-modality image registration has been presented. The experimental results show that PEM provides a more accurate and consistent registration performance between different modalities compared to the state-of-the-art methods. Therefore, we can conclude that PEM is more suitable for image-guided cardiac interventions where alignment of pre-operative images to inter-operative images is required. Moreover, PEM is a generic and modular image representation method; it can be applied to any other US, CT and MR image analysis problem.

3.4 Multi-Atlas Segmentation of Cardiac Ultrasound Images with PEMs

Automated left ventricle (LV) segmentation in 3D ultrasound (3D-US) remains a challenging research problem due to variable image quality and limited field-of-view. Modern segmentation approaches (shape, appearance and contour model based surface fitting) require an accurate initialization and good image boundary features to obtain reliable and consistent results. They are therefore not well suited for this problem. The method proposed in this section, overcomes

those limitations with a novel and generic 3D-US image boundary representation technique: Probabilistic Edge Map (PEM). This new representation captures regularized and complete edge responses from standard 3D-US images. PEM is utilized in a multi-atlas LV segmentation framework to spatially align target and atlas images. Experiments on data from the MICCAI 2014 CETUS challenge show that the proposed approach is better suited for LV segmentation than the active contour, appearance and voxel classification approaches, achieving lower surface distance errors and better LV volume estimates.

Clinical motivation: Cardiac ultrasound remains the primary imaging modality in the assessment of left ventricular systolic function, mass and volume to assess the morphology and function of the heart. Automated tools to analyse three-dimensional ultrasound (3D-US) images are important to ensure reproducibility as well as consistency of segmentations and to reduce the workload of clinicians. The development of such tools is still an ongoing research problem due to limitations posed by low image quality, restricted field-of-view and anatomical variations. For these reasons, accurate and generic image analysis techniques are crucial.

Related work: Automated left ventricle (LV) segmentation techniques can be broadly categorized into two groups: image-driven and model-driven approaches. Level-set approaches such as phase asymmetry [216] are part of the first category. They calculate 3D LV surfaces with weak or no shape constraints and do not require the fitting of a model to a large number of images. Also the B-spline active surface approach proposed in [14] does not require model fitting. Instead, the surface is initialized with an ellipsoid and B-splines are used to regularize the deformation of the surface model. Approaches in the second group use additional a-priori information by analyzing intensity patterns in training samples and manually traced contours. This includes approaches such as appearance models (AAM) [252] and semantic labelling of voxels using a classifier such as a decision forest [144]. Another method proposed in [193] uses labeled atlases and image registration to segment the LV volume. It does not require the training of a shape model, but makes an implicit use of such model through the atlases.

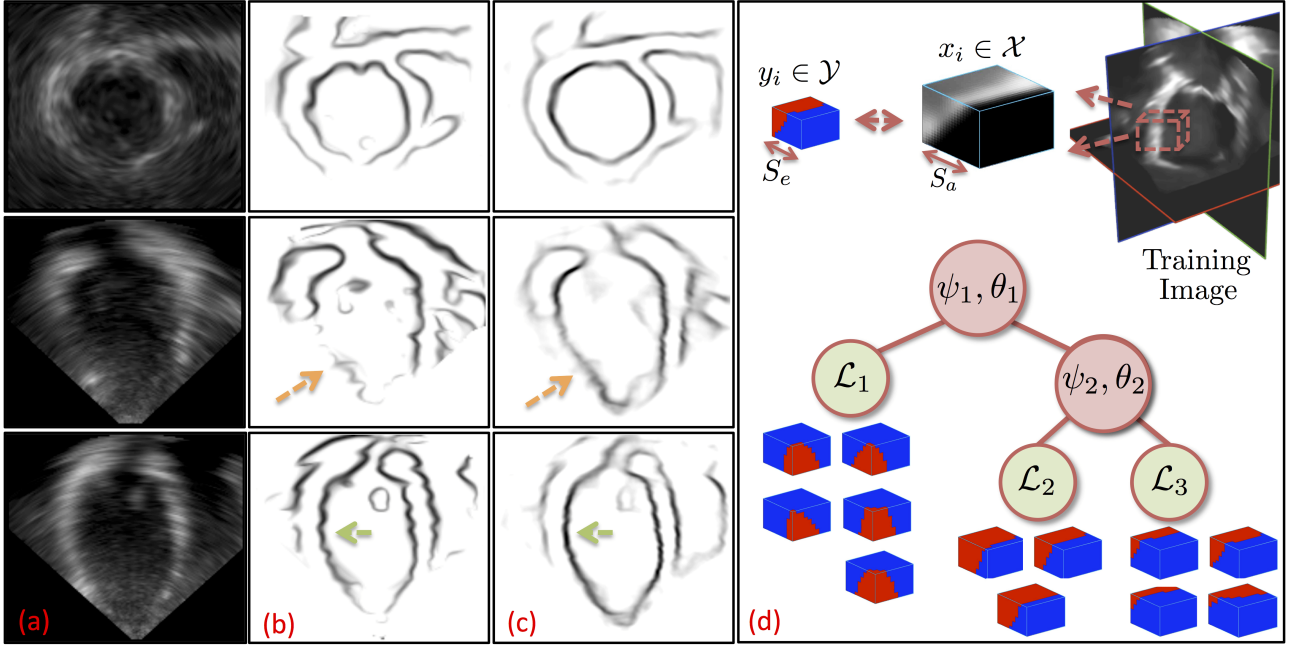


Figure 3.7: (a) 3D cardiac US image, (b) phase congruency [216], and (c) PEM which captures missing structures (orange arrows) and provides smoother edge response (green arrows). In (d) SDF training is illustrated, where the label patches (y_i) are clustered at each node split, and the weak learners (ψ_i) search for the optimal threshold value (θ_i) and feature (x_i) to separate the two clusters.

Active contour and level-set approaches require an accurate estimate of LV shape and position for initialization. This is because final segmentation results are sensitive to initializations obtained either manually [68, 193] or through ad-hoc solutions such as Hough transform of edges [14] or through selection of image center points [252]. Such approaches depend on the acquisition field-of-view and cannot be generalized to acquisitions from different acoustic windows such as apical and parasternal views together. Similarly, these approaches [14, 216, 252] make use of intensity and phase based features to delineate ventricle borders. Since phase features rely on the agreement of phases between different Fourier components (and are therefore insensitive to contrast), less importance is given to local energy information. This causes these features to be sensitive to noise. Likewise, intensity based approaches are sensitive to low image quality, shadowing, speckle and clutter.

Contributions: In this section, we propose a fully automatic multi-atlas LV segmentation framework for US images. Additionally, a novel robust 3D boundary representation method,

Probabilistic Edge Map (PEM), is presented and utilized within this framework to address the challenges outlined above. PEMs delineate object boundaries in the input images by using a trained structured decision forest (SDF) classifier [67]. With this method, we are extending the structural representation proposed in [68], applied on 2D cardiac short-axis slices, to a 3D structural analysis together with the use of US related image features. In this way, discontinuous and spurious edge responses in through plane direction can be eliminated, while achieving smooth and regularized tissue boundaries, as shown in Figure 3.7. In the proposed multi-atlas LV segmentation framework (PEM-MA), the PEMs are used in robust affine registration [198] and non-rigid registration [227] to spatially align multiple atlas images to the target. PEM based US image registration provides more reliable initialization between target and atlas images, and achieves better atlas selection [3] and LV segmentation performance. The proposed segmentation framework is evaluated on a benchmark dataset used in the MICCAI 2014 CETUS segmentation challenge. The results collected from the online evaluation platform show that PEM-MA achieves state-of-the-art LV segmentation accuracy in both surface distance and volumetric measure metrics, while outperforming all other challenge participants [13, 68, 252] in terms of the used evaluation criteria.

3.4.1 Methodology

Probabilistic edge map (PEM) representation: In cardiac imaging, 3D-US images outline an anatomical representation of the heart chambers. Further image analysis typically requires an accurate and smooth object boundary delineation. Data driven approaches may fail due to severe intensity artefacts and missing boundaries. A machine learning approach such as a structured decision forest (SDF) [67] can cope with these difficulties as the training data guides the boundary extraction. This is shown in Figure 3.7, where the proposed PEM captures the missing boundaries and delineates them accurately.

In the proposed segmentation framework, prior to the PEM generation, the US images are initially resampled to isotropic voxel size. Furthermore, speckle noise is reduced using a sparse

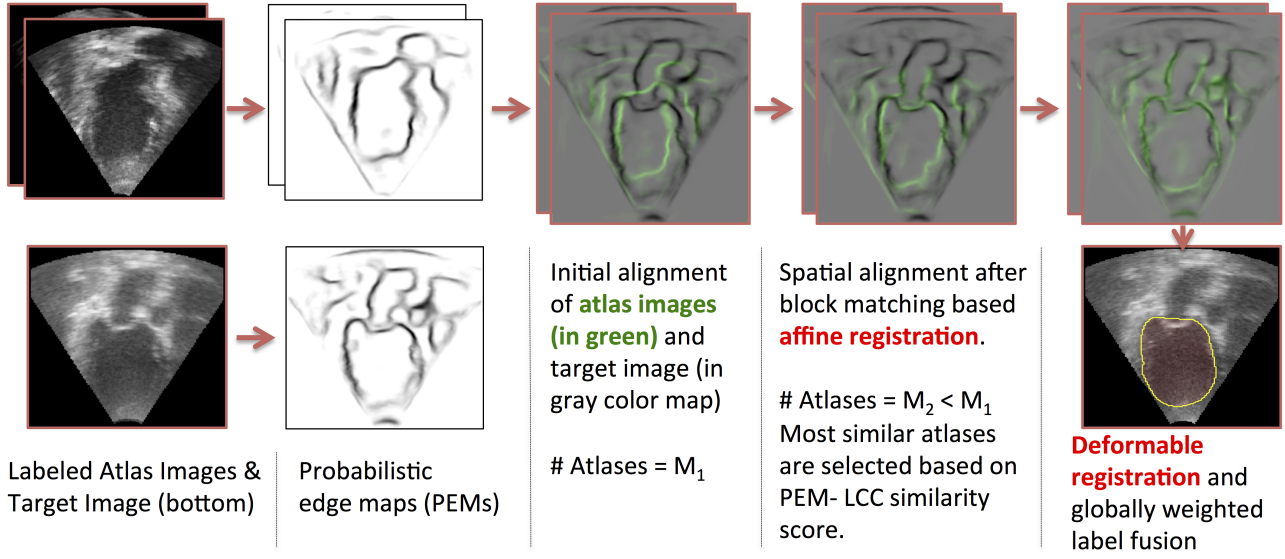


Figure 3.8: A block diagram of the proposed multi-atlas segmentation framework for 3D-US cardiac images.

coding approach: The K-SVD algorithm [75] is used to learn an over-complete dictionary from US image patches. After the learning stage, the image is reconstructed from a sparse combination of the learned dictionary atoms to remove speckle patterns. Finally, a SDF classifier for the PEM is trained from the preprocessed images. While SDFs are similar to decision forests, they possess several unique properties and advantages.

Multi-atlas left ventricle segmentation: Next, we detail our proposed multi-atlas LV segmentation framework as outlined in Figure 3.8, employing the generated edge maps. Initial affine alignment, atlas selection and deformable registration between target (I) and atlas images (J_i) are performed based on the PEMs (P^I, P_i^J) generated from the US images. A dataset consisting of a number of manually annotated US images is used in the atlas formation. The annotations for these atlases contain only the LV endocardial labels. The composite spatial transformations transfer the atlas labels to the target, followed by a globally weighted label fusion based on PEM similarity.

Global alignment: The PEMs from both target image and atlases are first aligned using a block matching technique [198] which maximizes the normalized correlation coefficient between

image blocks. The set of vectors defined by the displacement of each block is regularized before finding the global affine transformation A_i . A least trimmed squared regression based regularization (cf. [198]) removes the influence of displacements for the atlas blocks which have no target block correspondence due to missing features in the images. For this reason, this approach is robust to shadowing and anatomical variations and can provide an accurate spatial alignment for atlas selection and good initial segmentation.

Atlas selection: It was shown in multi-atlas brain segmentation [3], that a selection of most similar atlases is beneficial. Therefore, after affine registration, all M_1 atlases are ranked according to their average local correlation coefficient [36] score, $LCC(P^I, P_i^J \circ A_i)$, and the $M_2 < M_1$ top scoring atlases in the upper quartile are selected. The LCC similarity metric is defined in (3.3), where Ω denotes the target voxels within a region defined by the dilated LV mask.

$$LCC(P^I, P^J) = \frac{1}{|\Omega|} \sum_{x \in \Omega} \frac{|\langle P^I, P^J \rangle_x|}{\sqrt{\langle P^I, P^I \rangle_x \langle P^J, P^J \rangle_x}} \quad (3.3)$$

A Gaussian window G_σ with variance σ^2 locally weights the PEMs and $\langle P^I, P^J \rangle_x = G_\sigma * (P^I \cdot P^J)[x] - (G_\sigma * P^I)[x](G_\sigma * P^J)[x]$, where $\cdot$ denotes the Hadamard product, and $*$ the convolution. As the SDF classifier makes use of image intensities in node splits ψ , local intensity changes in the input images can influence the edge probabilities in PEMs. For this reason, LCC is a more suitable similarity measure for PEMs than global metrics such as sum of squared differences.

Local alignment: To correct for residual misalignment, a registration based on free-form deformations (FFDs) [227] follows the atlas selection. The total energy $E(T_i) = -LCC(P^I, P_i^J \circ T_i \circ A_i) + \lambda BE(T_i)$ is minimised in a multi-resolution scheme, where BE is the bending energy of the cubic B-spline FFD T_i and λ defines the trade-off between local PEM alignment and deformation smoothness.

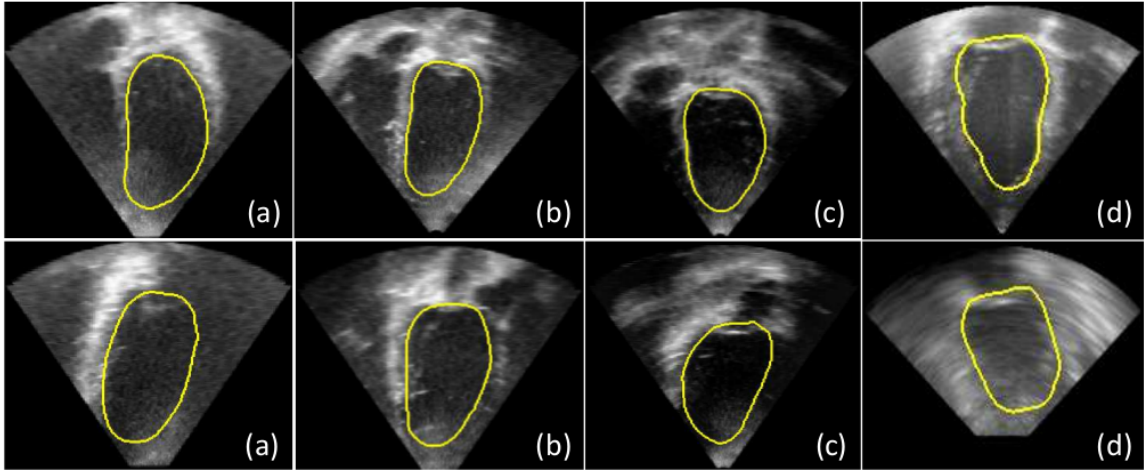


Figure 3.9: 3D-US cardiac image segmentation results obtained using multi-atlas fusion and PEM representation techniques. The images acquired from four different healthy subjects (a-d) are shown together with their automatically generated endocardial boundary contours. From left to right, best and worst segmentation results (in terms of Dice score) are shown.

Label fusion: Finally, the transferred atlas labels are fused using a globally weighted voting¹ [4] based on the dissimilarity $m_i = 1 - LCC(P^I, P_i^J \circ T_i \circ A_i)$. The LV segmentation of the target image is then given by the labelling function $S^I(x) = \arg \max_{l \in \{0,1\}} \sum_{i=1}^{M_2} w_i \cdot \delta(S_i^J(x) - l)$, where δ is the Dirac delta function and global weights $w_i = \exp(-m_i / \frac{1}{M_2} \sum_{j=1}^{M_2} m_j)$. In this fusion strategy, atlases more similar (higher LCC score) to the target image have a stronger influence on the final segmentation and those with a relatively lower score are downgraded.

3.4.2 Experiments and Results

The proposed segmentation framework is evaluated on a benchmark dataset used in the MICCAI 2014 CETUS challenge [199]. It consists of 4D echo sequences acquired from an apical window in healthy volunteers and patients with myocardial infarction and dilative cardiomyopathy. The dataset is divided into 15 training and 30 testing image sequences. Contours of the heart chambers were outlined by three experts, but only those of the training set are publicly available. Therefore, the CETUS website² is used for evaluation. Submissions are

¹Locally weighted and majority voting fusion methods were also evaluated in the experiments, and the best results were obtained with the global fusion method.

²<https://miccai.creatis.insa-lyon.fr/miccai/community/1>

Table 3.1: LV segmentation results on 30 subjects (CETUS challenge testing dataset [199]). Mean distance (MD [mm]), Hausdorff distance (HD [mm]) and Dice coefficient (DC [%]) results are listed separately for ED and ES frames. PEMs provide a better boundary representation than spectral features [193] based on mean ($p < 0.01$) and Hausdorff distances ($p < 0.01$)

	MD_{ED}	MD_{ES}	HD_{ED}	HD_{ES}	DC_{ED}	DC_{ES}
Manual [199]	1.01±.30	1.01±.38	3.37±.87	3.30±.94	0.949±.15	0.938±0.21
AAM [252]	2.44±.91	2.79±1.24	8.45±3.50	8.65±2.85	0.879±.05	0.835±.08
BEAS [13, 14]	2.26±.73	2.43±.91	8.10±2.66	8.13±3.08	0.894±.04	0.856±.06
SE-MA [193]	2.18±.70	2.47±.74	7.55±1.76	8.57±2.96	0.894±.03	0.849±.04
SDF-LS [68]	2.09±.68	2.20±.72	9.31±3.89	8.35±2.67	0.894±.04	0.871±.05
PEM-MA	1.94±.55	2.23±.60	7.00±1.99	7.53±2.23	0.904±.02	0.874±.04

automatically evaluated based on surface distance errors and clinical LV volumetric indices.

In all experiments, segmentations are computed only for end-diastolic (ED) and end-systolic (ES) phases. In Figure 3.9, PEM multi-atlas segmentation results obtained on images from four different subjects are shown. As can be seen in the figure, the multi-atlas technique allows us to obtain anatomically meaningful endocardial boundaries even if part of the left ventricle is missing in the images. Table 3.1 lists the surface distance errors obtained in the first experiment. The proposed PEM-MA framework achieves better results than the challenge top performing algorithms: AAM [252] (active appearance model), BEAS [13, 14] (B-spline active contours), SDF-LS (structured decision forest followed by level-set segmentation), and SE-MA [193] (spectral embedding multi-atlas method). The inter-observer manual segmentation [199] variations are reported for comparison. We can conclude that PEMs provide a better boundary representation than spectral features [193] based on mean ($p < 0.01$) and Hausdorff distance ($p < 0.01$). Moreover, the proposed approach does not require landmark selection [193] or manual affine alignment of LV surface template to initialize the segmentation [68].

The difference in segmentation accuracy between PEM-MA and model based surface fitting methods (AAM, BEAS) can be explained as follows. The proposed approach employs affinely aligned atlas labels as shape priors which are selected based on LCC similarity of PEMs, whereas the other methods use less data specific priors such as mean LV shape [252] and ellipsoid [14] shape assumption. Similarly, in PEM-MA, the LV segmentation is initialized with position

Table 3.2: Segmentation results on 30 images (CETUS testing data [199]). Pearson’s correlation coefficient (corr) and Bland-Altman ($\mu \pm 1.96\sigma$) limit of agreement (LOA) between ground-truth and estimated LV volume values are reported.

	ED_{corr}	ED_{LOA}	ES_{corr}	ES_{LOA}	EF_{corr}	EF_{LOA}
Manual [199]	0.981	-0.636 $\pm$ 18.2	0.987	-0.50 $\pm$ 14.4	0.959	0.13 $\pm$ 6.07
AAM [252]	0.966	-15.42 $\pm$ 32.1	0.964	-13.2 $\pm$ 28.9	0.611	3.69 $\pm$ 17.58
SE-MA [193]	0.945	-6.02 $\pm$ 41.6	0.924	-0.42 $\pm$ 41.2	0.780	-1.55 $\pm$ 13.88
BEAS [13, 14]	0.965	-4.99 $\pm$ 35.3	0.967	-6.78 $\pm$ 27.7	0.889	2.88 $\pm$ 10.48
SDF-LS [68]	0.917	8.73 $\pm$ 49.9	0.956	-5.16 $\pm$ 31.7	0.819	8.33 $\pm$ 14.46
PEM-MA	0.961	-4.14 $\pm$ 34.0	0.973	-3.47 $\pm$ 26.7	0.892	0.48 $\pm$ 10.78

priors obtained through a robust affine block matching of PEMs. This delineates the left ventricle position in the image more accurately than Hough transform [14] and the mean LV position of the training images [252].

In the second experiment, clinical indices, such as ejection fraction (EF), ED and ES volume values, are computed from the proposed segmentation approach. The obtained results are compared against their reference values using the aforementioned web site. The results in Table 3.2 show that PEM-MA achieves a better agreement with the ground truth compared to the other methods. As PEM-MA delineates LV boundaries more accurately, better volume estimates are obtained. Additionally, we observe that PEM-MA displays a consistent performance in both LV surface fitting and volume estimation in contrast to SDF-LS. The performance difference between the two can be linked to the improved structural representation and the choice of different surface fitting algorithm.

All experiments were carried out on a 3.00 GHz quad-core machine. The average computation time per image pair was 74s for non-rigid registration, 16s for affine alignment and 20s to compute each PEM. The training of the SDF (70m per tree) and atlas PEM computation were performed offline prior to target segmentations. The segmentation of the LV takes in total 16m per image. The proposed approach is computationally more complex than the methods in [14, 68] due to the multitude of registrations. However, a parallel implementation of these registrations significantly reduces the total runtime.

Implementation details: In total $N_t = 8$ PEM decision trees are trained using 20 US sequences plus rotated versions of these images. PEM quality was not improved further by including more trees. Patch sizes for training features and ground-truth edges are chosen as $S_a = 20$ and $S_e = 10$ per dimension. For global alignment, blocks of size 5^3 voxels were used with search radius equal to the block size as in [198]. A multi-scale optimization strategy was employed to capture large displacements and to improve convergence. A total of $M_1 = 30$ ED and ES atlases were aligned to each subject. Of these, on average $M_2 = 6.3$ were selected based on their LCC score, with a standard deviation of the Gaussian $\sigma = 7$ voxels in each dimension. The B-spline FFD control point spacing was set to 8 voxels per dimension for each image resolution. All transformations were computed using the NiftyReg library³ which provides an optimized implementation of the B-spline FFD algorithm [227].

3.4.3 Discussion and Conclusion

We presented a novel US image representation (PEM) which achieves state-of-the-art cardiac US image registration and LV segmentation results within a multi-atlas framework. The proposed framework outperforms all other methods participating in the MICCAI 2014 CETUS challenge based on the obtained surface mesh evaluation criteria. The main contributions of this section are: (1) highly accurate 3D edge map representation for cardiac US images, and (2) block-matching based robust and accurate initialization technique for automatic LV segmentation. The proposed PEM representation is generic and modular. It has the potential of being applied to echo images acquired from other organs and does not make assumptions on the acquisition window and image orientation. Additionally, the multi-atlas segmentation framework is shown to be applicable for clinical routine as it can estimate functional indices very accurately.

³<http://sourceforge.net/projects/niftyreg/>

3.5 PEM Based Respiratory Motion Correction in Cardiac Cine-MR Images

2D short axis (SAX) cine MR images are of crucial importance for the accurate assessment of cardiac anatomy and function. Since SAX cine stacks are routinely acquired during multiple breath-holds, different breath-hold positions can cause a misalignment of the heart between different slices, with potential detrimental effects on a variety of clinically relevant measurements (e.g. volume or shape of the left ventricle). In this section, we propose a novel approach to spatially align motion corrupted SA slices in MR image stacks using 3D probabilistic edge maps (PEMs) generated with structured decision forests. In our technique, each 2D SAX slice is associated with a 3D PEM outlining the myocardial contours in the same slice as well as in the adjacent one. In-plane spatial misalignment between adjacent slices is then corrected using a registration algorithm applied to the associated PEMs. This approach was tested against a conventional intensity-based registration method on SAX cine stacks acquired from 26 healthy subjects, for whom anatomical 3D cardiac images were also available as reference. End-diastolic left ventricular volumes are estimated using a 3D multi-atlas segmentation technique and used to quantify alignment accuracy. The results show that the proposed technique successfully reduces the misalignment between slices and that the registered stacks allow a more accurate volumetric estimation than both the original and the intensity-corrected ones.

Clinical motivation: Cardiovascular magnetic resonance (CMR) imagery can be used for accurate volumetric reconstruction of the beating heart throughout the cardiac cycle, which is in turn necessary for a great number of clinically-relevant tasks such as volumetric estimations and myocardial motion analysis. While fast SSFP sequences allow nowadays the direct acquisition of an anatomical 3D image (A3D) of the whole heart, they are usually limited by either relatively poor image quality or low temporal resolution, making them unsuitable for accurate functional assessment. Consequently, the most common CMR sequence currently used in the clinical practice is still the short axis (SAX) SSFP cine, consisting of 10-14 parallel slices and 20-30

frames per cardiac cycle. SAX cine stacks are generated during multiple breath-holds (i.e. 1-3 slices acquired per each breath-hold). Although the subjects are instructed to hold their breath at the same breath-hold position, in practice the heart location can vary considerably. If the differences between the breath-hold positions are not accounted for with suitable inter-slice alignment procedures, the acquired image stack will not correctly represent the cardiac volume, introducing potential errors in all of the following analyses and visualizations.

Related work: Several post-processing techniques for respiratory motion correction for SAX cine images have been presented in the last years. Virtually all of the proposed approaches consist in the 3D rigid registration between the acquired SAX cine stack and other CMR images, such as A3D [42, 292] or long axis (LAX) [22, 37] images. Unfortunately, in the clinical practice these additional sequences are often unavailable. Therefore, inter-slice alignment can only be performed by conventional intensity-based approaches, which are on the other hand relatively inaccurate. The reason for this mainly lies in the considerable changes in the imaged features between adjacent slices (due to the relatively high slice thickness of a typical SAX cine stack), which hinders the application of registration approaches based on the comparison of information directly extracted from the two slices.

In this chapter, we propose a novel approach to in-plane spatial alignment of motion corrupted SAX slices in MR image stacks using 3D probabilistic edge maps (PEMs) generated with structured decision forests. PEMs are learning-based image representations outlining the contours of a specific object of interest (the myocardium, in this case). Through training, structured decision forests are able to associate to each 2D SAX image for slice i a 3D PEM representing the contours at slices i and $i+1$. The overlap between 3D PEMs generated from adjacent slices allows to correct the in-plane spatial misalignment using a block matching registration algorithm. The proposed approach was tested on a dataset of real SAX cine stacks against a conventional intensity-based registration method.

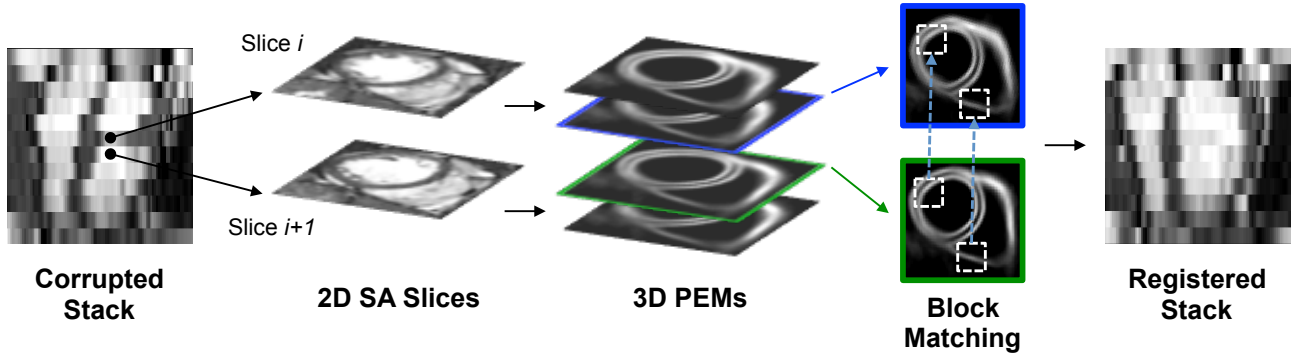


Figure 3.10: Diagram of the proposed motion correction technique: 3D PEMs are estimated from each pair of adjacent SAX slices and then aligned using a block matching algorithm. The estimated transformation is then applied to the actual slice.

3.5.1 Methodology

In our technique, the described approach to edge representation was applied to associate to each 2D SAX image for slice i a 3D PEM outlining the myocardial contours at slices i and $i+1$ (see Figure 3.10). The PEM at slice $i-1$ could be also predicted, but this would be at the expense of the prediction accuracy as the dimensionality of output space would be increased. To build a structured decision forest model with motion-free data, the training dataset consisted of slices extracted from the A3D acquisitions (downsampled in the z direction to match the slice thickness of the 2D cine data later used for testing) together with myocardial contours (the edge maps used as labels). During testing, the structured random forest predicts for each 2D SAX image at slice i not only the myocardial contours location for the same slice but also for the following one ($i+1$) in a motion-free scenario. The overlap between 3D PEMs estimated for adjacent slices of the stack allows to apply a block matching registration algorithm [197] to perform rigid in-plane spatial re-alignment. The rigid transformation estimated for the PEMs is then applied to the actual image at slice $i+1$. Each slice (from second to last) undergoes the same process, resulting in the complete re-alignment of the stack from base to apex.

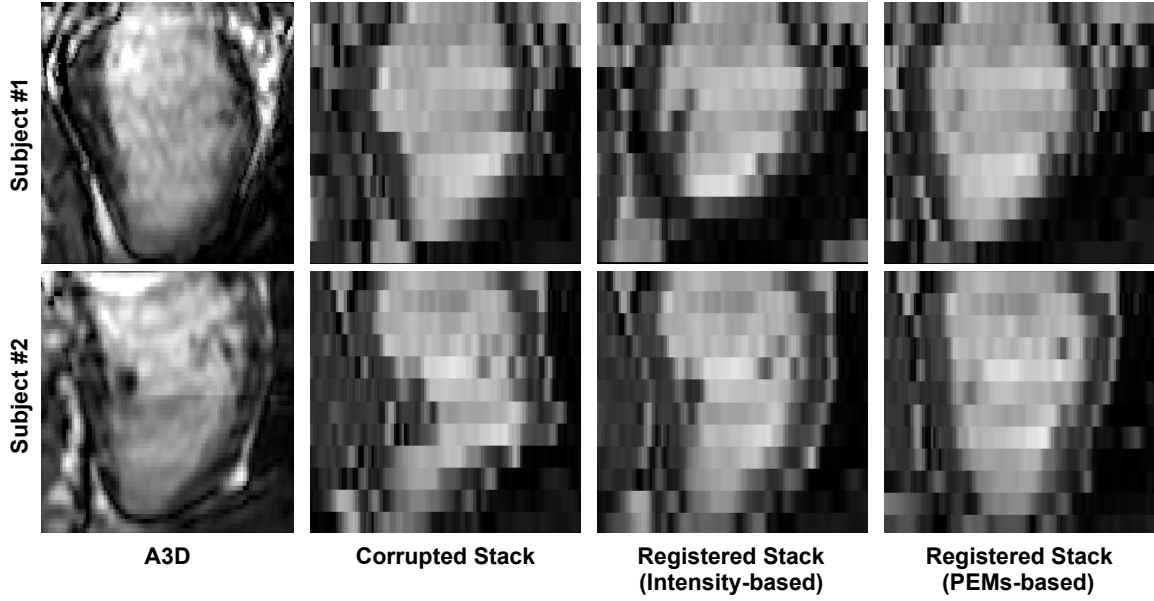


Figure 3.11: Respiratory motion correction results obtained in two different subjects using the proposed technique (PEMs-based) and a conventional intensity-based registration approach. 3D high-resolution (A3D) images (first column) and 2D SAX stacks (second column) are also displayed for comparison.

3.5.2 Experiments and Results

Imaging data: CMR imaging was performed on 435 healthy subjects using a 1.5T Philips Achieva system (Best, Netherlands) equipped with a 32 element cardiac phased-array coil. 2D cine balanced steady-state free precession (b-SSFP) images were acquired in the left ventricular short axis (LVSA) plane from base to apex using the following parameters: repetition time/echo time 3.0/1.5 ms, flip angle 60° , acquired pixel size 2.0×2.2 mm, slice thickness 8 mm (2 mm gap); reconstructed voxel size $1.2 \times 1.2 \times 8$ mm; number of slices 10-12, cardiac phases 30. A single breath-hold A3D LVSA b-SSFP sequence was also acquired in the same orientation.

Experimental details and results: The structured decision forest model was trained on 410 A3D images at end diastole - together with the associated myocardial contours - following the present implementation details: image patch size 36×36 px, edge map labels size $12 \times 12 \times 2$ px, number of trees $T = 16$.

The proposed registration technique was tested on 25 2D cine stacks at the end diastolic phase.

Table 3.3: End-diastolic volume measurement differences (ΔV) between different respiratory motion correction approaches: intensity (int) and PEM based image registration approaches. The computed volumes are compared against the ground-truth values obtained from the high-resolution 3D images. The accuracy of the volumetric measurements prior to the alignments is shown on the first line.

Methods	ΔV (mL)	p -vals
$ \text{EDV}_{\text{pre}} - \text{EDV}_{\text{ref}} $	10.34 ± 9.41	-
$ \text{EDV}_{\text{int}} - \text{EDV}_{\text{ref}} $	10.76 ± 9.64	-
$ \text{EDV}_{\text{pem}} - \text{EDV}_{\text{ref}} $	7.73 ± 6.55	$p = .016$

For comparison, we also tested a conventional intensity-based registration technique, in which the block matching algorithm [197] was directly applied to the image intensity values between adjacent slices. The block matching algorithm was selected as competitor because in our experiments it consistently outperformed the more common methods based on normalized mutual information. A 3D multi-atlas segmentation technique [11] was applied to estimate end-diastolic left ventricular volumes respectively from the A3D images (EDV_{ref}), the 2D SAX image stacks before (EDV_{pre}) and after slice-alignment using the proposed technique (EDV_{pem}) and using the conventional intensity-based registration method (EDV_{int}). Since the motion-free A3D images and 2D SAX image stacks were acquired from the same subjects, EDV_{ref} could be used as reference, and volume differences between EDV_{ref} and EDV_{pre} , EDV_{pem} and EDV_{int} , respectively, could be used as proxy to accuracy of motion correction algorithms. Figure 3.11 shows examples of obtained registration results: qualitatively the proposed registration technique (PEMs-based) does a better job in slice realignment compared to the conventional (intensity-based) one. Statistics obtained from the computed volume differences are reported in Table 3.3: volume estimation performed on stacks registered using the proposed technique is more reliable in comparison to the conventional technique.

3.5.3 Discussion and Conclusion

In this section, we presented a novel respiratory motion correction technique that spatially aligns motion-corrupted image slices in 2D SAX MR stacks using 3D PEMs. The proposed

approach has been successfully applied to 3D multi-modal registration between MR and either ultrasound or CT images [189]. Differently from that work, in which 3D PEMs were associated with 3D image patches, the structured decision forest presented in this section associates 3D PEMs to 2D image patches extracted from a single image slice. The obtained results show that the proposed approach was able to perform slice realignment with higher accuracy than the conventional intensity-based technique (as far as volume estimation was concerned). Importantly, the proposed approach is currently limited to in-plane spatial alignment. Out-of-plane motion correction from a single image stack is in fact an ill-posed problem, and while the availability of other sequences (e.g. other SAX or LAX stacks) would allow to perform registration also in the z direction, these data may not always be available.

In conclusion, the proposed technique provides higher registration accuracy for the spatial alignment of 2D SAX slices in comparison to a conventional intensity-based method, and therefore, when no additional sequences are available, it should be adopted as a pre-processing step to volumetric estimation from cine CMR image stacks.

3.6 Other Applications of PEMs

Besides the common application areas presented (e.g. registration, segmentation) in the previous sections, PEM representation can also be utilised in clinical studies. One particular example is the comparison of cardiac imaging modalities, such as ultrasound (US) and MRI, in terms of quantitative measurements of ventricular volume. In clinical studies [170, 241], it has been reported that US based volume measurements tend to be significantly lower than the values obtained from cardiac MRI which is considered to be the gold-standard in clinical routine. However, a detailed analysis on the underlying causes of this difference has not been investigated in depth. In particular, we are interested in identifying the LV regions where the images acquired with these modalities show differences.

For the analysis, we make use of the clinical dataset provided in the MICCAI'11 STACOM

Challenge [256], which contains 15 pairs of US/MR cardiac images acquired from healthy adult subjects. The image processing pipeline to analyse these images can be described as follows: (I) each US and MR image is automatically segmented using the multi-atlas technique described in Section 3.4. (II) The images and their corresponding endocardium segmentations are rigidly aligned using the PEM registration method (cf. Section 3.3). (III) To compute the surface distances between the spatially aligned US and MR segmentations, the PEIS similarity metric [140] is used. PEIS can be considered as a more informative metric in comparison to the standard surface distances since it looks for similar segmentation patterns. The described framework is illustrated in Figure 3.12. The distance map, visualised for one subject in Figure

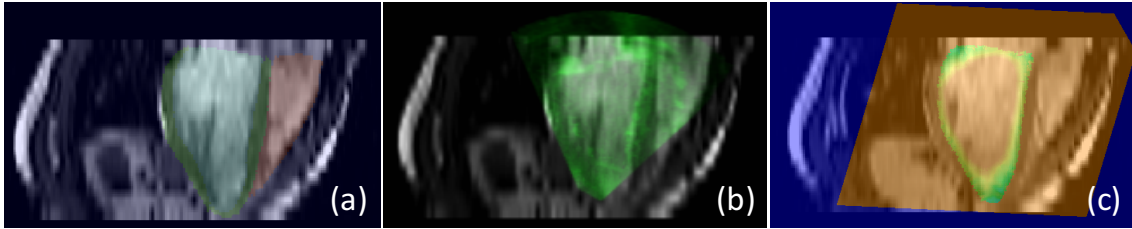


Figure 3.12: The proposed image analysis framework to compare cardiac US and MR imaging modalities. (a) Images from each modality are segmented using a multi-atlas technique. (b) The multi-modal images are registered using the PEM representation, where the US image is shown in green colour-map. Later, the spatially aligned endocardial segmentations are compared with each other using the PEIS metric [140]. (c) The surface distances (PEIS) are shown in green colormap, the errors are clustered around the apical and basal parts of the left ventricle.

3.12 (c), is computed also for the remaining US/MR pairs. These results are then mapped to the cardiac left-ventricle atlas [10] to aggregate the results from all the US/MR pairs and obtain a mean representation of the distance errors, as can be seen in Figure 3.13. The cardiac

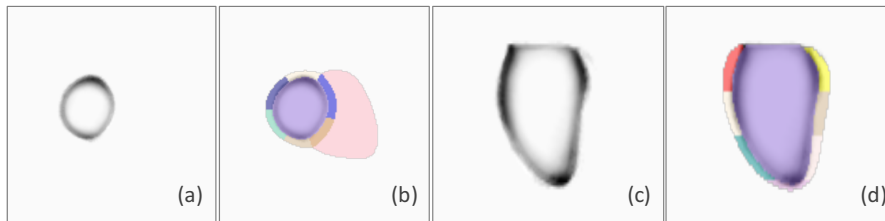


Figure 3.13: Surface distance errors obtained from US/MR endocardium segmentations, are averaged after mapping them to the cardiac atlas space. The AHA-17 atlas model (b,d) is used to identify LV regions where the errors are accumulated. In particular, we see that the differences are observed in the anterior, apical and basal parts of the LV.

atlas, shown in (b-d), is represented with the AHA-17 segment model [41]. As can be seen in Figure 3.13, the surface errors are mainly concentrated around the apical and basal slices. More importantly, as reported in a similar study [99], segmentations also show large differences around the anterior wall. We can interpret this results with the fact that the anterior wall is often not visible in US images due to shadowing from the air in the lung. For this reason, the segmentations performed on these images may not match the underlying true anatomical structure, whereas it is visible in cardiac MR images.

As another application area of PEM representation, we have analysed its applicability to images acquired from other organs of interest, such as abdominal CT images. In that regard, a structured decision forest is trained with 150 CT images and liver boundary annotations. The results on the testing set, shown in Figure 3.14, demonstrate that PEM multi-atlas and registration techniques can be extended to abdominal image analysis as well.

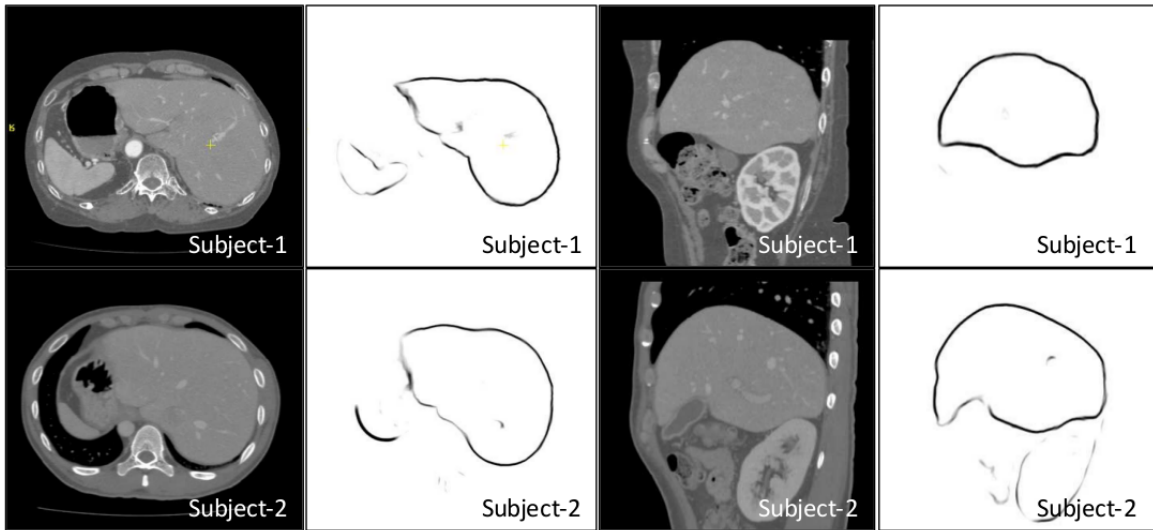


Figure 3.14: Probabilistic edge maps generated from the abdominal CT images. The extracted contours outline the surface of the liver.

3.7 Conclusion

In this chapter, we have presented a novel cardiac image representation technique, namely Probabilistic Edge Maps (PEMs), that can be used to bridge the gap between different imaging

modalities. By mapping multi-modal images to the same PEM space, we are able to identify the correspondences between these images. In this way, a more accurate similarity metric can be defined between the modalities such as ultrasound and MRI. The applicability of this new representation type to cardiac image analysis problems has been demonstrated with the experiments presented in this chapter. The application areas reported on are: (I) multi-modal cardiac image registration, (II) multi-atlas ultrasound image segmentation, and (III) cardiac cine-MR respiratory motion correction. Moreover, in the last section, we have qualitatively demonstrated that the proposed technique can additionally be used to extract the surface representation of the abdominal organs.

Recent work [274] has shown that probabilistic surface representations can also be extracted through convolutional neural networks (CNNs). Our initial experiments on the HED model [274] have shown that CNNs are able to learn more distinctive features and suppress false positive detections, which are usually observed in the background when PEMs are extracted with SDFs. However, CNN based structured learning has not been studied in depth and we believe that future research could investigate the use of PEM-CNN models to further improve the cardiac image analysis results. Follow-up work could also investigate 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 structural surface representation is experimentally shown to be a reasonable approach to solve this problem, 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 PEM representation by introducing a new decision tree model, namely stratified decision forests. The new approach enables us to learn population specific regression and classification models, which results in improved accuracy in anatomical landmark localisation.

Chapter 4

Stratified Decision Forest: Applied to Anatomical Landmark Localisation

This chapter is based on:

O. Oktay, W. Bai, R. Guerrero, M. Rajchl, A. de Marvao, D. P. O'Regan, S. Cook, M. P. Heinrich, B. Glocker, and D. Rueckert, “Stratified Decision Forests for Accurate Anatomical Landmark Localisation in Cardiac Images”, *IEEE Transactions on Medical Imaging*, 36(1), pp. 332-42, 2017.

Abstract: Accurate localisation of anatomical landmarks is an important step in medical imaging, as it provides useful prior information for subsequent image analysis and acquisition methods. It is particularly useful for initialization of automatic image analysis tools (e.g. segmentation and registration) and detection of scan planes for automated image acquisition. Landmark localisation has been commonly performed using learning based approaches, such as classifier and/or regressor models. However, trained models may not generalize well in heterogeneous datasets when the images contain large differences due to size, pose and shape variations of organs. To learn more data-adaptive and patient specific models, we propose a

novel stratification based training model, and demonstrate its use in a decision forest. The proposed approach does not require any additional training information compared to the standard model training procedure and can be easily integrated into any decision tree framework. The proposed method is evaluated on 1080 3D high-resolution and 90 multi-stack 2D cardiac cine MR images. The experiments show that the proposed method achieves state-of-the-art landmark localisation accuracy and outperforms standard regression and classification based approaches. Additionally, the proposed method is used in a multi-atlas segmentation to create a fully automatic segmentation pipeline, and the results show that it achieves state-of-the-art segmentation accuracy.

4.1 Introduction

Accurate detection of anatomical landmarks is important for clinical applications that require fully-automated image segmentation and registration. In many cases, landmark localisation is a prerequisite for the initialization of subsequent image analysis steps, such as initialization of deformable model and atlas based approaches [11] for cardiac modelling. Similarly, detected landmarks can be used to facilitate fully automatic planning of image acquisitions, such as cardiac MRI examinations [162]. Additionally, landmark localisation in cardiac images, e.g. right ventricle (RV) insertion points, can be used to analyse left ventricle function according to AHA myocardial segmentation models [41]. However, variable imaging quality and variations of the heart's shape, size and orientation across subjects and populations pose a great challenge to fully automatically detect landmarks from medical images, particularly in cardiac MR images.

A common approach to locate landmarks of interest is based on predictions of a trained classifier or regressor. This achieves reliable and accurate results as long as both the learned model and the training data are similar enough. However, existing approaches mostly use averages or other simple statistics over subsets of training examples which can yield large prediction errors on test data. In this chapter, we propose a generic learning approach (stratification) that can

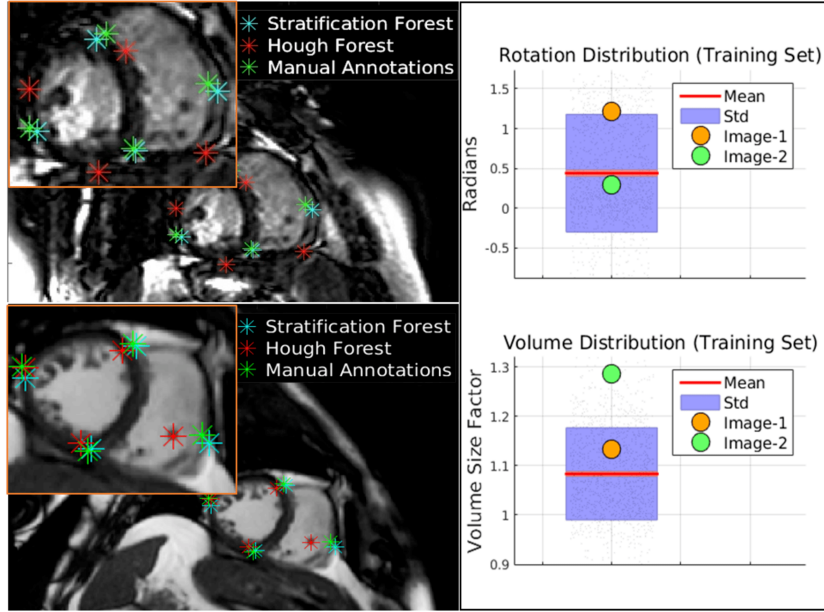


Figure 4.1: A comparison between standard Hough and proposed stratified decision forests for landmark localisation in two different short axis cardiac MR images. The first image (top) has a significantly different heart orientation than the population average shown by the red bar (right); similarly the heart in the second image (bottom) is larger in terms of size in comparison to the average heart size in the training images. The proposed approach is able to cope with these variations by learning adaptive and patient-specific regression models.

be used in the context of decision trees to learn more adaptive classifiers and regressors without requiring additional training information. In that regard, we extend the SDF model presented in the previous chapter by including stratification node splits within the tree structure. With the proposed method, models are trained not only with local information collected from image patches, but also with global information such as shape, size, and pose differences of the organs. Indeed, it is the main motivation for the use of the term stratification for the proposed method. In this way, the landmark detection accuracy can be improved as shown in Figure 4.1: This example shows that the proposed method can localize the anatomical landmarks accurately although the organ of interest (in this case the heart) exhibits large variations in terms of pose, size and shape. The next section discusses the existing work on landmark and object localisation approaches proposed in computer vision and medical imaging.

4.1.1 Related work

Landmark localisation has been a particular research area in both computer vision and medical image analysis over the last decades. Landmark localisation methods in both medical and natural images can be mainly categorized into two subgroups, namely image registration and model based approaches. The approaches in the former category identify dense correspondences between target and atlas images to localize anatomical landmarks. Robust image alignment [9, 197, 262] and non-rigid registration [53, 227] techniques can be used to relate anatomical correspondences between images.

The approaches in the later category, on the other hand, learn discriminative and generative models through training datasets to associate anatomical landmark locations with prior information, such as shape and appearance information of tissues. These approaches can be further categorised into three subgroups as classification, regression, and graphical models [209]. In addition to these categories, there are several examples such as Hough Forests [90] in which the regression and classification approaches are combined in a hybrid manner. In the first category, work based on the marginal space learning (MSL) framework [290] proposed the use of cascaded boosting tree detectors to identify pose and landmark locations in medical images. The presented cascaded classifiers method [160, 162] is applied on 2D cardiac MR images to detect landmarks. In computer vision, Shotton et al. [245] used trained classifier trees to identify body parts in depth images, which is used as an intermediate step to locate joint positions in human body. Similarly, Donner et al. [73] proposed a decision forest based classifier to locate hand joint locations in X-ray images. More recent work [289] proposed the use of separable filters that are learned via multi-layer perceptrons, in boosting tree classifiers. However, the main limitation of classifier based approaches is caused by the limited number of positive training samples. Moreover, classifier based approaches are susceptible to imaging artefacts and low imaging quality in image regions where the landmarks are located. In ultrasound images, shadowing [183] can be another limitation for the classifier based approaches, which is similar to occlusions in natural images.

To avoid these problems, regression based methods combine the knowledge from different image regions that can be near or distant from the target landmark locations. Aggregation of predictions from several regions can yield more robust localisation and addresses the problems mentioned above. Particularly, for the skeletal joint localisation problem, recent methods [95, 179] proposed regression based localisation approaches and showed an improved performance over the classification methods. Similarly, in the context of medical image analysis, Criminisi et al. [56] have used regression trees to locate 3D bounding boxes corresponding to organs in abdominal CT images. In a recent work [91], the regression formulation is modified by changing the standard offset model parameters with atlas scale and position regression parameters. This was shown to achieve improved localisation performance in CT images as it makes use of global shape information of organs.

Hybrid approaches, on the other hand, integrate regression and classification models in a single detector framework. These approaches benefit from richer annotations as the regression models are conditioned on segmentation labels. In [236], a joint model is proposed to locate vehicles in natural images. Similarly, Hough forests [90] are used for pedestrian and horse detection in natural images.

The common parametric model used in regression models is based on multivariate normal distributions. For this reason, the learned models may not be specific enough for all cases, in particular where the sub-populations form multiple clusters, e.g. different heart size and shape in medical images. Similarly, in the context of facial image analysis, the pose of the human head is an important latent variable that can affect the model performance. For facial gesture recognition, the methods presented in [61, 286] address this problem by proposing conditional and hierarchical decision trees. However, the conditional trees require additional classifiers to estimate the probability of this latent variable, and based on its value the required decision trees are selected for inference. Similarly, spectral forest methods [153] have been proposed to allow population specific bagging to train specialized decision trees, which showed an improved segmentation performance compared to the standard bagging. However, the proposed tree

selection process in testing requires the computation of nearest neighbours for each test image as a pre-processing step.

4.1.2 Contributions

In this chapter, we present a single unified decision tree training approach that generalizes the previously presented cascaded localisation frameworks. Particularly, the latent variable parameters are computed within the stratified trees in a probabilistic way rather than using auto-context models as in MSL. Additionally, the proposed approach does not require any additional training information, specialized decision trees [61], or externally applied dimensionality reduction methods [153] for training and inference. Our method is along the same line as the previously proposed neighbourhood approximation forests [130], where anatomically similar training images are grouped together in the sub-tree architecture to obtain population clusters. Differently, in the proposed method, we show that this type of clustering with stratification nodes can enhance the classification and regression performance without adding significant computational costs. The proposed novelties can be characterized as follows:

- A stratification training objective is presented to learn more data adaptive decision tree models in an implicit way. In this way, the learned model can be more representative for datasets showing large variations such as object shape and size.
- The proposed model does not require cascaded classifiers [160] or specialized trees [61, 153] to learn the statistics of latent variables, such as pose parameters. Therefore, it can be considered as a generalization of cascaded models and can be more easily adapted to other classification/regression problems.
- The presented work shows the application of joint classification and regression based approach on anatomical landmark localisation in cardiac MR images.

Besides the proposed novelties, the method presented in this chapter is built on some existing

Table 4.1: A short summary of related work, decision forest based classification and regression approaches in the literature

Related Work	Method Description
Criminisi et al. [56]	A regression forest based landmark detection method is applied to locate visceral organs in full-body CT images.
Gall et al. [90]	Hough forest method is presented for joint training of classification and regression nodes.
Konukoglu et al. [130]	Neighbourhood approximation forest (NAF) is proposed to cluster brain MR images based on subjects' age and pathology group. In testing time, the learned models are used to find most similar subjects among the training images.

classification and regression techniques in the literature (see Table 4.1). The improvements upon these existing approaches can be described as follows:

- Criminisi et al. [56]: Regression splits are defined in a similar way, but the leaf node prediction models are characterised by the prior information which is the probability of class labels and global shape characteristics.
- Gall et al. [90]: In addition to classification and regression labels, global characteristics of organs (e.g. shape/size/pose) are incorporated in a tree model to learn more representative regression models.
- Konukoglu et al. [130]: We show that clustering of training samples based on their global characteristics, such as size and shape, can actually improve classification and regression accuracy. In contrast to this, [130] is used in subject clustering and nearest-subject search problems.

The proposed method is validated on 1080 3D and 90 multi-stack 2D short-axis cardiac MR images acquired with different scanners. The results show that it achieves more accurate localisation results compared to Hough forests and MSL based cascaded classifier localisation

methods. The experiments also show that it provides better initialization for subsequent multi atlas image segmentation [207].

4.2 Methodology

A decision tree is a tree-structured predictor consisting of two types of nodes, split nodes $\psi \in \mathcal{T}$ and leaf nodes $\ell \in \mathcal{L}$. Split nodes route samples $x \in \mathcal{X}$ to leaf nodes to find the best match for a given sample against a set of training examples, whereas leaf nodes store the posterior distributions $p(y|x)$ for output variable $y \in \mathcal{Y}$ and make a prediction for the sample $x \in \mathcal{X}$. Split nodes are characterized by decision stumps $\psi(x) : \mathcal{X} \rightarrow \{0, 1\}$ which route samples to left and right sub-decision trees. An ensemble of uncorrelated decision tree classifiers is known as a decision forest [31].

Here we propose a new stratification training objective to learn node split models that can group samples based on their global characteristics such as organ shape, size, and orientation. The presented training scheme includes structured classification, regression, and stratification split nodes, which will be explained in the following sections.

4.2.1 Model Input Space

The input space of the decision trees is characterized by image channels $\mathbf{x} \in \mathcal{X}$, and there are C channels for each training sample $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^C)$. The channels are defined by multi-resolution appearance, histograms of gradients (HoG) [60], and gradient magnitude image patches of size $(M_a)^3$. The appearance channels are formed by constructing a two-scale Gaussian image pyramid, where the original input is downsampled by factor of two in the top layer. The smoothed gradient magnitude and orientation are computed with oriented Gaussian derivative kernels. Orientation histograms are computed using soft-binning, where bin weights are determined by the gradient magnitude (cf. [66]). The features for each patch $\mathcal{P}_i$ centred at

$\mathbf{p}_c$ are extracted from each channel $\alpha \in \{1, 2, \dots, C\}$ by performing comparisons of the average intensity of boxes (R_1, R_2) within the patch boundaries in a similar fashion to [59, 245].

$$q_\lambda(\mathbf{p}_c) = \frac{1}{|R_1|} \sum_{\mathbf{p} \in R_1} x^\alpha(\mathbf{p}) - \frac{1}{|R_2|} \sum_{\mathbf{p} \in R_2} x^\alpha(\mathbf{p}) \quad (4.1)$$

For a single dimensional decision stump, the split node parameters are defined by the parameter set $\lambda = (R_1, R_2, \alpha)$ and threshold value γ . For pairwise channel comparisons, the box regions R_i are chosen to be non-zero, and for single channel comparisons R_2 is chosen to be zero. Using the same notation, the split function is defined as:

$$\psi(q_\lambda(\mathbf{p}_c)) = \begin{cases} 1, & \text{if } q_\lambda(\cdot) \geq \gamma \\ 0, & \text{otherwise} \end{cases} \quad (4.2)$$

4.2.2 Structured Classification

In the proposed method, the information stored in the leaf nodes (output space $\mathcal{Y}$) is characterized by class labels, regression, and stratification model parameters. In other words, a single decision tree model is learned to perform multiple tasks simultaneously, such as organ surface delineation, landmark location regression, and shape information regression. The joint training of classification and regression nodes benefit from each other to learn more representative and class specific models since more ground-truth information is provided during training as suggested by Gall et al. [90].

For the organ surface delineation task, the class labels are stored in the leaf nodes as label patches by applying the structured learning methods proposed in [67, 129] which have been shown to increase the segmentation accuracy over single point estimates as they produce regularized and smoother class labellings. Dollar et al. [67] proposed a method to enable the split of training samples $\mathcal{P}_i = (x_i(\mathbf{p}), y_i(\mathbf{p}))$ with label patches $y \in \mathbb{Z}^{(M_e)^3}$ of size $(M_e)^3$. The label patches are mapped into an intermediate space in which the Euclidean distances between the

samples can be used to perform dimensionality reduction. In this way, the label patches are mapped into one dimensional space and the standard entropy based training objective $\mathcal{H}_c$ can be used to split the training samples.

In a related work [189], this approach is applied on 3D cardiac MR and ultrasound images to generate probabilistic surface maps of the heart in medical images. Similarly, in this work, the leaf nodes cast predictions to find the probabilistic map of the object surface $E(\mathbf{p})$ for the given input channels.

$$p(E(\mathbf{p}) | \mathbf{x}) = \frac{1}{|\Omega_{\mathbf{p}}| \cdot N_{\text{tree}}} \sum_{i \in \Omega_{\mathbf{p}}} \sum_{\ell \in \mathcal{L}_i} p(y(\mathbf{p}_i) = 1 | \ell) \quad (4.3)$$

where $\Omega_{\mathbf{p}}$ is the set of input patches that are overlapping at the location $\mathbf{p}$. N_{tree} denotes the number of trees and $p(y(\mathbf{p}_i) = 1 | \ell)$ is the probability of the event whether the binary edge patch stored in the leaf node ℓ is positive at $\mathbf{p}$ or not. The learned class posterior distributions in the leaf nodes are used as weighting terms in the regression function. The regression node training and inference information is provided in the next section.

4.2.3 Regression of Landmark Locations

The structured forest model can be enhanced by adding regression nodes in addition to the structured classification split nodes. In this way, the trained classifier can cast regression votes for anatomical landmark locations. More importantly, this combination enables to train class specific regression models. Each training sample $\{\mathcal{P}_i = (x_i, y_i, \mathcal{D}_i)\}$ is now characterized by a set of offset vectors to each landmark point. The set of offsets $\mathcal{D}_i = (\mathbf{d}_i^1, \mathbf{d}_i^2, \dots, \mathbf{d}_i^L)$ contain 3D displacement vectors from the patch centre to the target landmark location, where L is the number of landmarks.

The regression split nodes are trained by minimizing the determinant of full covariance matrix [56] defined by the landmark offset vectors. In this way, the inter-dependency between the

landmark locations are partially taken into account, and it allows the model to learn an implicit shape model of the organ.

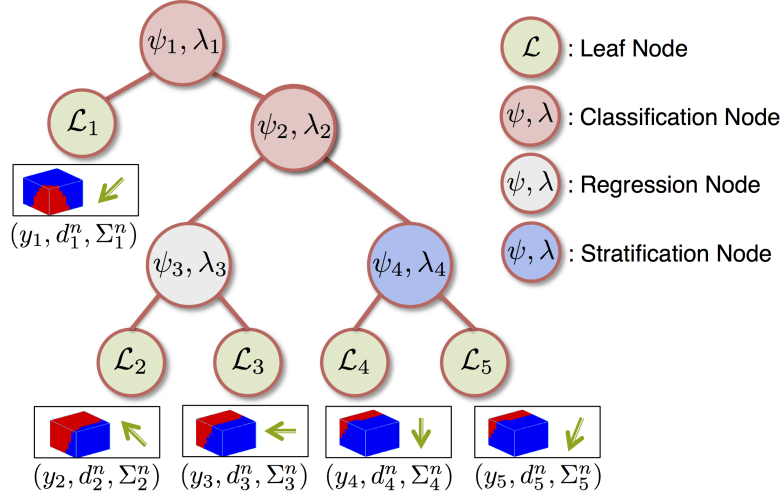


Figure 4.2: An illustration of the stratified decision tree structure. The standard Hough tree structure [90] is modified by adding the proposed stratification node splits.

In the leaf nodes ℓ , the regression information from the training samples is stored using a parametric model $p(\mathbf{d}_\ell^n) = \mathcal{N}(\mathbf{d}_\ell^n; \bar{\mathbf{d}}_\ell^n, \Sigma_\ell^n)$: The mean $\bar{\mathbf{d}}_\ell^n$ and covariance matrix Σ_ℓ^n parameters are computed for the $3L$ dimensional multivariate normal distribution. This regression model is preferred over the non-parametric models, such as Parzen estimation or Mean-shift mode seeking, due to its computational simplicity. Similar to the work proposed in [236], offset vector models in the leaf nodes are conditioned based on segmentation label of the training samples. In other words, samples collected around the target organ surface have a separate regression model $p(\mathbf{d}^n | y = k, \ell)$ than the background samples. As in [90, 236], we assume that the background pixels are not as informative as the organ surface voxels in detecting anatomical landmarks. As the pose variations can be quite large, long range regression votes are observed to be decreasing the landmark localisation accuracy and confidence. Therefore, the Hough vote maps $F(\mathbf{p}^n)$ are formulated as follows:

$$\begin{aligned}
 p(F(\mathbf{p}^n) | \mathbf{x}) &= \sum_{k=\{0,1\}} p(F(\mathbf{p}^n), y(\mathbf{p}) = k | \mathbf{x}) \\
 p(F(\mathbf{p}^n) | \mathbf{x}) &\approx p(F(\mathbf{p}^n) | \mathbf{x}, y(\mathbf{p}) = 1) p(y(\mathbf{p}) = 1)
 \end{aligned} \tag{4.4}$$

Here only the points along the organ surface ($y(\mathbf{p}) = 1$) are allowed to cast votes for the landmark locations. However, landmarks can be positioned at any arbitrary location either inside or along the organ surface. The probability of a voxel classified as a point on the organ surface is obtained through the use of classification node splits (Section 4.2.2). In this way, the landmark prediction function is constructed as:

$$p(F(\mathbf{p}^n) | \mathbf{x}) \approx \frac{1}{N_{\text{tree}} \cdot N} \sum_{i=1}^N \sum_{\ell \in \mathcal{L}_i} w_{\ell}^n p(E(\mathbf{p} | \mathbf{x})) K \left(\frac{\mathbf{p}^n - (\mathbf{p}_{(i,c)} + \overline{\mathbf{d}}_{\ell}^n)}{h} \right) \quad (4.5)$$

where N is the number of image voxels, K is a Gaussian kernel with bandwidth parameter h , $w_{\ell}^n = \frac{1}{\text{trace}(\Sigma_{\ell}^n)}$ is the confidence parameter for landmark n . The subscript (i, c) denotes the centre voxel of input patch. After computing the Hough vote map, the landmark location is determined by choosing the voxel with highest probability value as proposed in [56]:

$$\hat{\mathbf{p}}^n = \arg \max_{\mathbf{p}^n} p(F(\mathbf{p}^n) | \mathbf{x}) \quad (4.6)$$

4.2.4 Stratification of Global Characteristics

Modelling the offset vectors to landmarks in leaf nodes as Gaussian distribution biases the learning towards the average landmark distribution observed in training data. In datasets with smaller variations in terms of pose and object shape, this is unlikely to introduce too much error. However, in cases such as cardiac MR imaging, the orientation and size of the heart can exhibit large degrees of variation. For these cases, it is useful to have population groupings in sub-trees to increase the localisation accuracy.

Along the same line, neighbourhood approximation forests [130] were proposed to cluster brain MR images of patients from different population groups. Our method takes this approach one step further by making use of the implicit clustering in a decision tree composed of classification and regression nodes. This clustering of the data can be viewed as a population stratification

and allows our method to achieve improved landmark localisation accuracy. The proposed method does not require additional training data and information compared to standard Hough forests.

To achieve this goal, the training images are annotated with size and pose parameters. These parameters are computed automatically using only the given input landmark point sets. A reference point set $\mathcal{F}_r = \{\mathbf{p}_r^1, \mathbf{p}_r^2, \dots, \mathbf{p}_r^L\}$, obtained from a selected reference MR cardiac image, is used to quantify in-plane orientation θ_i and size β_i of a given training image. To obtain these parameters for each training image i , the ground-truth landmark point set $\mathcal{F}_i$ is aligned to $\mathcal{F}_r$ by computing an affine transformation $A_{(i,r)}$.

The meta information $\mathcal{M}_i = \{\theta_i, \beta_i\}$ is used in stratification nodes ψ_s to cluster training samples with similar pose and size. The impurity of training samples $\mathcal{S} = \{\mathcal{P}_i = (x_i, y_i, \mathcal{D}_i, \mathcal{M}_i)\}$ in terms of the size parameter is defined based on mean squared differences $H_\beta(\mathcal{S}) = \sum_i (\beta_i - \tilde{\beta})^2$. Similarly we compute the impurity in terms of the orientation parameter $H_\theta(\cdot)$. The two impurity measures H_β and H_θ may have quite different ranges depending on the problem and its dimensionality, and one of them could easily dominate the other one during optimization. Hence, we combine the two in a single stratification uncertainty term $H_s(\cdot)$ by normalizing the two uncertainty terms similar to the joint training objective proposed in [97]:

$$H_s(\mathcal{S}) = \frac{1}{2} \left(\frac{H_\theta(\mathcal{S})}{H_\theta(\mathcal{S}_0)} + \frac{H_\beta(\mathcal{S})}{H_\beta(\mathcal{S}_0)} \right) \quad (4.7)$$

Here $\mathcal{S}_0$ denotes the training sample set reaching the root node, and $\mathcal{S}$ represents the selected samples after the node split. With this formulation, we assume that the random variables corresponding to size and orientation are independent of each other, i.e. the joint entropy can be expressed as $p(\theta, \beta | \mathbf{x}) = p(\theta | \mathbf{x}) \cdot p(\beta | \mathbf{x})$. In the negative log domain, the joint entropy is expressed as summation of the two terms, which leads to the equation in (4.7).

As shown in Figure 4.2, the trees are trained in a joint manner using the following split node types: structured classification ψ_c , regression ψ_r , and stratification nodes ψ_s . As proposed

in [90], the training objective at each split node is selected randomly among the listed three objectives. Based on our initial experiments, the node selection probabilities are fixed to $p(c) = p(r) = 0.4$, $p(s) = 0.2$.

The stratification splits can be trained with the box features presented in Section 4.2.1. However, global scale features are found to be more expressive in separating images in terms of shape and pose parameters as shown in the work [286] of iterative facial landmark localisation. For this reason, we introduce global shape features to represent all samples collected from the same image, and they are used only for the stratification splits. These features are (i) inter-landmark distances and distance ratios [286], which are calculated through initial landmark location predictions, and (ii) histograms of gradients (HoG) [60] computed using only the initial edge map prediction of the target organ. Based on our experiments, these features achieve better stratification and regression results; however, their computation requires a two stage cascaded model as the landmark distance features are computed based on the initial estimates of the landmark locations. In other words, the initial tree model in the cascaded approach is used only to obtain a coarse representation of the organ surface and landmark locations, which are later used to perform stratification splits in training of subsequent decision trees.

4.2.5 Visualization of the Stratification Splits

To better understand the advantages of the proposed method, a proximity analysis is performed on a trained stratified decision forest. In the training procedure, this analysis is normally not required to be done; however, it is a useful technique to understand the role of the stratification splits. In more detail, we are able to visualize the mapping of the training images from the root node to leaf nodes. The computed proximities will demonstrate that images with similar organ size and pose parameters are automatically mapped to closer leaf nodes in the tree structure.

To perform this analysis, the trained forest is analysed by computing the proximity matrix of the training samples, as explained in [31], which is a $M \times M$ matrix and M is the total number

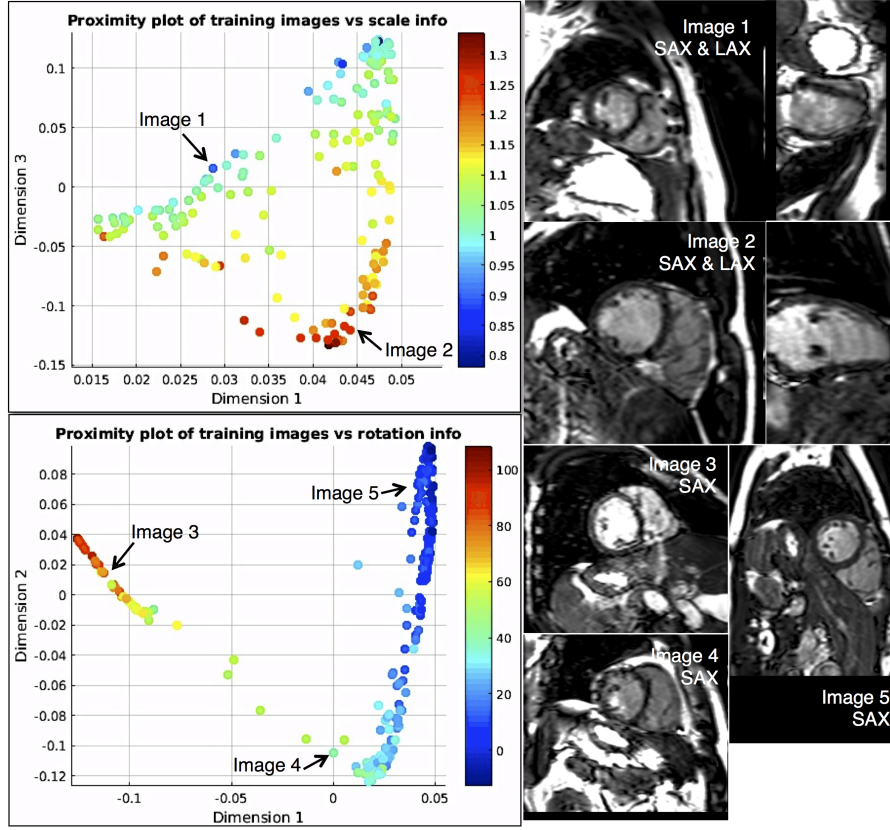


Figure 4.3: Proximity plot (left) of the training images is shown, where the color code displays the heart orientation (bottom) and size (top) information of each image. The axes correspond to the first 3 dimensions in the lower dimensional space. The selected images are visualized on the right side to demonstrate the size and pose variations.

of training images. If the two images are mapped into the same leaf node, then their proximity is increased by one. Similarly, an adjacency matrix can be derived from these proximities, and the connections between the images can be visualized by applying a non-linear dimensionality reduction technique. In our setting, the adjacency matrix is analysed using the Laplacian Eigenmaps method [17], and the images are mapped into a two dimensional manifold space as can be seen in Figure 4.3.

In this figure, it can be seen that the cardiac short axis images with similar in-plane orientation and heart size are mapped into closer leaf nodes and share the same sub-trees. In this way, regression and classification information stored in leaf nodes are population specific, and it allows the nodes to make more accurate landmark predictions.

4.3 Experiments and Results

For the training and evaluation of the proposed method, we used two separate and disjoint cardiac MR datasets, which are referred to as *Dataset1* and *Dataset2*. The first set contains 1080 cine short-axis MR images of resolution 1.50x1.50x2.00 mm from the UK Digital Heart Project [10]. *Dataset1* is randomly partitioned into three equal sized subsets for 3-fold cross-validation, so in total 720 images are used to train the models for all the methods. The second dataset consists of 90 cine short-axis MR images with lower resolution 1.50x1.50x8.00 mm. This dataset is part of the Cardiac Atlas Project (CAP) [83] and it is publicly available. The images from both datasets were acquired in different clinical sites with different MR scanners. As a preprocessing step, all images are linearly up-sampled to the same resolution as the images in the first dataset. Additionally, a data-augmentation strategy [132] is utilized to increase the number of training samples, which is performed by using label-preserving spatial transformations.

For evaluation of the proposed method, two different types of experiments are performed using these datasets. The first experiment demonstrates the accuracy of different landmark localisation methods. The proposed stratified forest is compared against the standard Hough forest and image registration based localisation techniques in two separate sub-experiments. For all the methods, we used the same training datasets and data-augmentation strategy. In the first experiment, the proposed method is compared to the baseline localisation results obtained from the standard Hough forest. In the second experiment, image registration methods (block-matching [197] and 3D-SIFT alignment [262]) are employed prior to Hough forest to reduce the pose and size variations of the heart observed in the training and testing images. The spatially aligned images are later processed with Hough forest for training and inference purposes.

The second evaluation focuses on the application of the proposed method for image segmentation. A state-of-the-art multi-atlas segmentation method [207] is augmented with the proposed landmark localisation method, and the obtained results are compared against the current semi-automatic approach in which the landmarks are identified manually.

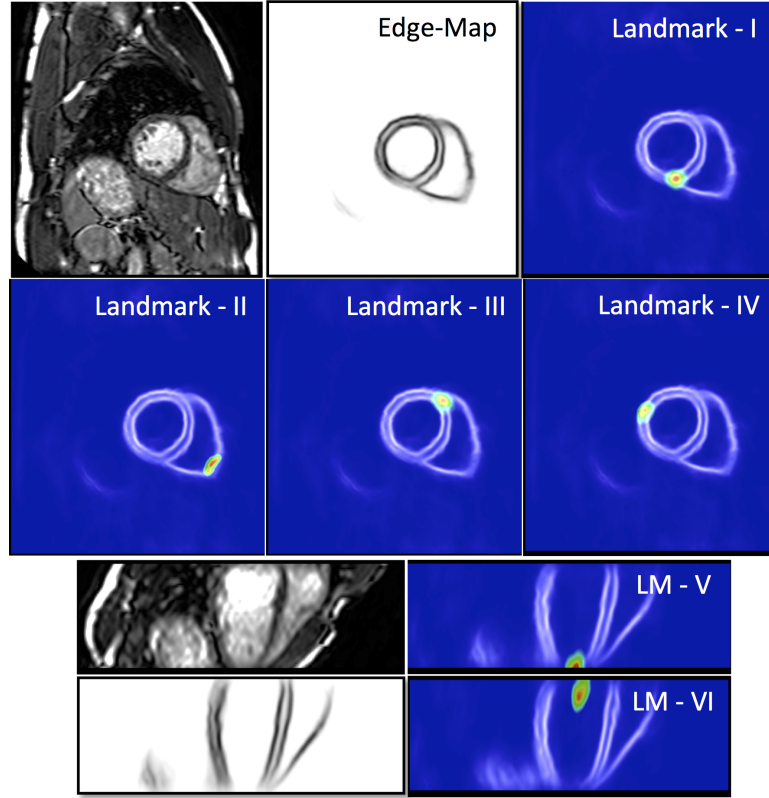


Figure 4.4: From top left to bottom right: Input cardiac MR image, probabilistic surface map of the heart, and Hough vote maps for the location of six different anatomical landmarks. The Hough vote maps, shown in jet color-map, are obtained with the stratified forest, and they are overlaid on top of the probabilistic surface maps. Voxels with high probability are shown in red color.

4.3.1 Anatomical Landmark Localisation

The decision trees for Hough and stratified forests are trained to locate $|L| = 6$ anatomical landmark locations. These landmark locations correspond to LV lateral wall mid-point, RV insert points (intersection between the RV outer boundary and the LV epicardium), RV lateral wall turning point (the point where the RV outer boundary changes directions significantly within the image), apex, and center of the mitral valve. The anatomical landmarks are shown in Figure 4.4. In Table 4.2, the detection results for these landmark locations are shown, which are produced by averaging the results obtained from the three folds of the cross-validation. The localisation errors are reported in terms of mean and median of Euclidean distance between the detected landmark position and the corresponding ground truth which is manually annotated by two experts. Additionally, the location of the center point of these landmarks $\hat{\mathbf{p}}^c = \frac{1}{L} \sum_{n=1}^L \hat{\mathbf{p}}^n$

Table 4.2: Landmark localisation errors for the two datasets, namely *Dataset1* (1.5x1.5x2.00 mm) and *Dataset2* (1.5x1.5x8.0 mm). The landmark localisation errors are reported in terms of mean and median Euclidean distances for all 6 landmark locations. Additionally, the localisation error for the centre point of all the landmarks is provided in the central column. The last column shows the localisation errors in each orthogonal direction in the image space. The proposed stratified forest (D) is compared against the 3D-SIFT robust alignment [262] & Hough forest (A), standard Hough forest (B), and block matching [197] & Hough forest (C) approaches. Please refer to Section 4.3 for a detailed description of the approaches. The inter-observer errors are reported in (E).

		Individual Landmarks, $ L = 6$		Centre Point of Landmarks	
		Mean (mm)	Med (mm)	Mean (mm)	Med (mm)
<i>Dataset1</i>	(A)	14.51±10.74	12.236	6.92±5.33	5.187
	(B)	10.74±6.32	9.224	6.05±4.12	5.572
	(C)	10.44±8.24	8.399	5.33±4.30	4.283
	(D)	8.03±5.11	7.067	4.10±2.81	3.422
	(E)	6.67±4.17	6.293	2.90±3.42	2.842
<i>Dataset2</i>	(A)	16.36±10.35	14.798	8.48±6.82	6.928
	(B)	12.42±6.12	11.245	7.36±3.61	6.314
	(C)	12.54±10.25	10.587	7.10±6.36	5.385
	(D)	9.12±4.64	8.412	4.62±2.63	4.312
	(E)	6.82±3.95	6.197	2.96±2.04	2.719

Distribution of the Errors				
		X (mm)	Y (mm)	Z(mm)
<i>Dataset1</i>	(A)	6.53±6.83	6.47±6.62	7.70±7.64
	(B)	5.36±4.38	5.47±4.45	5.86±4.89
	(C)	4.21±4.92	4.46±5.14	5.22±6.41
	(D)	3.37±3.78	3.31±3.73	3.65±4.13
	(E)	2.44±3.61	2.40±3.94	2.99±5.50
<i>Dataset2</i>	(A)	7.72±8.09	6.83±7.09	8.07±6.65
	(B)	5.75±4.23	5.62±4.23	7.65±5.64
	(C)	5.86±8.04	5.19±5.20	6.20±6.99
	(D)	3.65±3.31	3.51±2.13	5.43±5.71
	(E)	2.54±2.74	2.57±3.42	3.03±3.16

is also computed as this point is less influenced by the inter-observer variability in the manual annotations. The experimental results show that the use of the proposed stratification split nodes significantly increases the landmark localisation accuracy as it is better able to cope with variations of the size and pose of the heart as observed in the cardiac MR images. Table 4.2 also

shows the distribution of the errors along each image axis. The errors are mostly concentrated in the through plane direction due to the lower resolution in that direction. In Table 4.3, the localisation errors for each landmark point are shown.

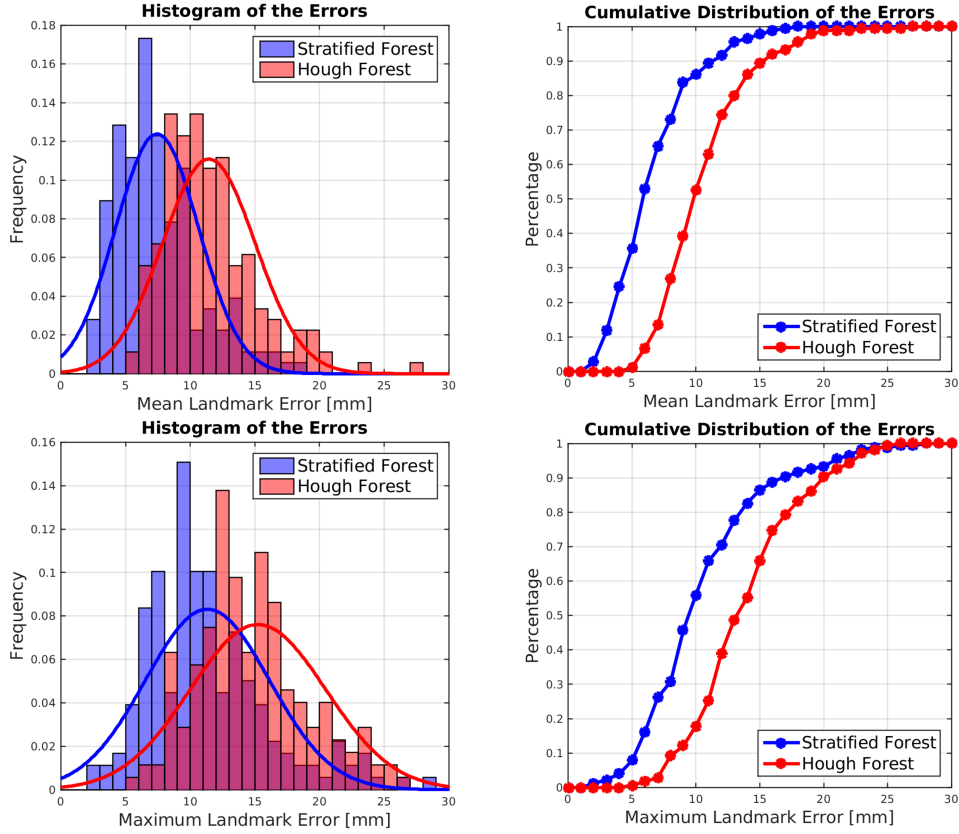


Figure 4.5: Histogram of the landmark localisation errors for the *Dataset1*. The distribution of mean (top) and maximum (bottom) localisation errors are shown.

The predictions for the second landmark (RV lateral) have a higher error compared to the other landmarks. The significant performance difference is due to lack of consistent definition of RV lateral landmark, and the large shape variability of RV wall. These factors also increase the variance in manual annotations for this particular landmark. Compared to the Hough forest based localisation, the proposed approach improves the detection accuracy significantly in both of the datasets. To observe the number of failure cases, the distribution of the errors is shown in the histogram in Figure 4.5. One can see that there are only a few outliers in the histogram. Moreover, the proposed method consistently performs better than the Hough forest based method as it can be seen on the cumulative distribution of the errors.

Table 4.3: Landmark localisation errors for the two datasets, namely *Dataset1* (1.5x1.5x2.00 mm) and *Dataset2* (1.5x1.5x8.0 mm). The localisation errors for each landmark point is provided in terms of mean Euclidean distance. The proposed stratified forest (D) is compared against the 3D-SIFT robust alignment [262] & Hough forest (A), standard Hough forest (B), and block matching [197] & Hough forest (C) approaches. The relative performance improvement (C-D) is shown in the last row (p-val $\ll 1e-3$). Please refer to Section 4.3 for a detailed description of the approaches. The inter-observer errors are reported in (E).

		LM1 (mm)	LM2 (mm)	LM3 (mm)	LM4 (mm)	LM5 (mm)	LM6 (mm)
<i>Dataset1</i>	(A)	11.14 $\pm$ 7.62	21.90 $\pm$ 13.84	13.00 $\pm$ 9.71	13.64 $\pm$ 7.68	16.49 $\pm$ 11.76	10.94 $\pm$ 7.88
	(B)	9.71 $\pm$ 5.65	17.03 $\pm$ 8.88	10.65 $\pm$ 5.20	10.18 $\pm$ 5.77	8.62 $\pm$ 4.55	8.27 $\pm$ 4.51
	(C)	8.80 $\pm$ 6.73	18.00 $\pm$ 16.73	9.51 $\pm$ 6.75	10.62 $\pm$ 8.26	7.99 $\pm$ 5.71	7.70 $\pm$ 5.25
	(D)	7.02 $\pm$ 4.93	12.23 $\pm$ 6.94	7.92 $\pm$ 5.52	7.36 $\pm$ 4.22	6.74 $\pm$ 4.12	6.32 $\pm$ 3.95
	(E)	5.88 $\pm$ 4.65	10.60 $\pm$ 8.09	5.60 $\pm$ 5.79	6.67 $\pm$ 5.30	5.97 $\pm$ 3.28	5.30 $\pm$ 2.98
	(C-D)	1.78	5.77	1.59	3.26	1.25	1.38
<i>Dataset2</i>	(A)	13.28 $\pm$ 8.60	21.96 $\pm$ 9.70	12.95 $\pm$ 7.11	15.92 $\pm$ 8.75	19.06 $\pm$ 8.53	15.02 $\pm$ 13.47
	(B)	12.12 $\pm$ 6.87	19.03 $\pm$ 12.31	10.41 $\pm$ 6.64	12.31 $\pm$ 5.94	11.31 $\pm$ 4.21	9.13 $\pm$ 4.53
	(C)	11.78 $\pm$ 11.04	18.76 $\pm$ 19.76	10.16 $\pm$ 6.83	13.14 $\pm$ 12.18	12.27 $\pm$ 6.30	9.15 $\pm$ 5.47
	(D)	8.42 $\pm$ 4.42	10.12 $\pm$ 5.02	8.04 $\pm$ 4.08	11.32 $\pm$ 4.16	10.41 $\pm$ 4.15	7.21 $\pm$ 3.38
	(E)	6.99 $\pm$ 4.40	8.34 $\pm$ 4.52	6.78 $\pm$ 4.37	6.66 $\pm$ 3.91	6.23 $\pm$ 2.89	5.96 $\pm$ 3.61
	(C-D)	3.36	8.64	2.12	1.82	1.86	1.93

It is observed that a slight performance improvement in mean accuracy can be achieved when the Hough forest is preceded by a robust image alignment method. However, this improvement comes at the cost of increased variance of errors when the registration algorithm fails to align images. This situation can be explained by two reasons: (i) Variation in the training data is reduced due to spatial alignment of images to a reference space. (ii) Image alignment results are susceptible to large anatomical variations observed in organs other than heart, which leads to incorrect alignment results as the heart label information is not used. In particular, the poor registration results obtained with 3D SIFT features are attributed to the large slice thickness, which reduces the number of matched features between the images.

In addition to the landmark localisation results, the performance of the pose and size estimations is measured. The root mean square (RMS) is adopted to evaluate the pose estimation results, and the ground truth information is obtained by globally aligning the manually annotated landmark points with an affine transformation. The pose estimation results for both datasets are shown in Table 4.4. The results show that the proposed stratified forest estimates the target

Table 4.4: Heart pose and size estimation accuracy results for the *Dataset1* (*DS1*) and *Dataset2* (*DS2*). Distribution of the ground-truth pose values are reported in terms of value range and standard deviation. The rotation values are given in radians and scale values are unitless.

Dataset and pose parameter		Ground-truth value range	Standard deviation	Estimation error (RMS)
<i>DS1</i>	Rotation	[-0.24 1.81]	0.501	0.11±0.08
	Size	[0.86 1.33]	0.097	0.05±0.03
<i>DS2</i>	Rotation	[-0.36 1.67]	0.447	0.12±0.11
	Size	[0.91 1.32]	0.083	0.05±0.04

pose and size parameters very accurately, which results in improved landmark localisation performance.

4.3.2 Cardiac MR Image Segmentation

Landmark localisation can be useful for subsequent image analysis such as segmentation and registration. In our experiments, a state-of-the-art multi-atlas patch based segmentation technique [54] is selected for this task. In comparison to classifier based methods, multi-atlas approaches have been more successful in the semantic segmentation of cardiac MR images as reported in the recent RV segmentation challenge [207]; however, they often require manual initializations. For this reason, the proposed landmark localisation is used in a multi-atlas segmentation framework to create a fully-automatic segmentation pipeline.

The localized landmarks are used to initialize the registration algorithm between the target and atlas images. A similarity transformation with 9 degrees of freedom is computed using the detected landmark point sets. The global alignment is followed by a B-spline based free-form deformable image registration [227]. The propagated labels are fused together using majority voting, and a graph-cut segmentation is applied as a post-processing to fill the gaps and smooth the segmentation labels.

In the experiments, four different segmentation frameworks are tested. All the frameworks made use of the same multi-atlas segmentation method, but the registration initializations

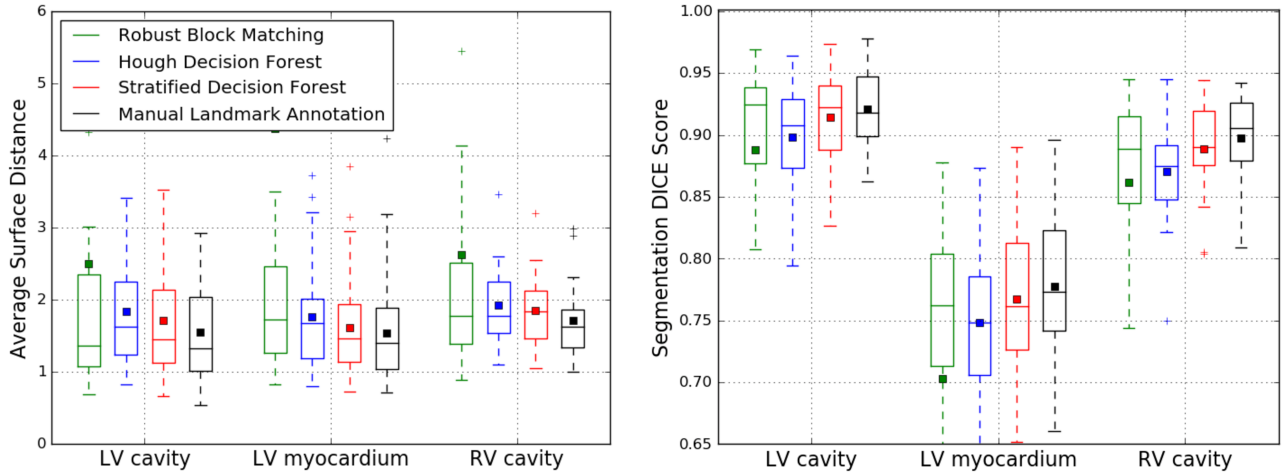


Figure 4.6: Cardiac MR image multi-atlas segmentation results: dice score and mean surface distances are obtained for four different registration initialization techniques. Mean values are shown with coloured square boxes.

were done either with different landmark point sets or by using an affine image registration. In that respect, we used manual annotations (I), Hough forest based landmarks (II), the proposed stratified decision forest based landmarks (III), and a robust block matching method (IV) [197]. In this experiment, we investigate the influence of landmark localisation accuracy on segmentation results. Also, we compare the performance difference between the landmark and affine registration based initialization [293] on the segmentation method. Here, a separate dataset of cardiac images is used, which is publicly available and part of the Cardiac Atlas Project [83]. The dataset was used in the MICCAI'13 SATA challenge to benchmark different cardiac segmentation methods. It consists of 50 manually annotated image sequences acquired from healthy subjects, patients with LV hypertrophy and wall abnormalities due to prior myocardial infarction. Segmentations are performed only on the end-diastolic frames extracted from the sequences, and RV segmentation labels are collected by manually annotating the images. The segmentation results, shown in Figure 4.6, are achieved using 20 fixed atlases selected from the dataset.

Based on the results, we observe that the robust block matching method (RBM), in some cases, completely fails to find correspondences between the atlas and target images, which causes the observed outliers and large variations in the segmentation accuracy. It can be mainly explained

with orientation and shape differences between the target and atlas images, and RBM is more sensitive to these conditions. On the other hand, the results show that most of the landmark localisation errors can be compensated during the segmentation process. For this reason, the performance difference between the two automatic localisation methods is less prominent in the segmentation results. However, we observe that in some cases the Hough forest landmark predictions fail, and this results in incorrect segmentations as can be seen in Figure 4.6 in form of outliers. More importantly, it is observed that the proposed landmark localisation and segmentation initialization method yields similar accuracy as the semi-automatic segmentation method that is based on manual landmark identification. Therefore, we conclude that the proposed method is accurate enough in landmark localisation to guide a multi-atlas segmentation method.

In a different experimental setting, the stratified forest and block matching methods were applied together sequentially to initialize the multi-atlas segmentation. This approach did not improve the segmentation accuracy significantly. However, as the computational cost increase significantly with block matching, in practice one can safely opt for a stratified forest only based initialization.

4.3.3 Implementation Details

In Table 4.5, the parameter values used in the experiments are specified. Better localisation results could be obtained on the same datasets by tuning the parameter values, which was not performed in our experiments. As the leaf nodes store structured patch label information, the number of required randomized trees is not as large as the standard decision forest approaches. Landmark localisation and block matching experiments were carried out on a Intel-i7 3.40 GHz quad-core machine, and the approximate computation time to initialize each atlas was 12s for stratified forest and 2.1min for block matching. The non-rigid registration experiments were conducted on a machine with a graphical processing unit (GPU), and the average computation time was 49.2s per atlas. Therefore, with the proposed fully-automatic segmentation pipeline

Table 4.5: Stratification forest parameter specifications

Parameter value	Description
$(M_a)^3 = (28)^3$	3D Input patch size
$(M_e)^3 = (12)^3$	3D Output PEM patch size
$N_{\text{tree}} = 10$	Number of decision trees
$L = 6$	Number of landmarks
$D_{\text{max}} = 36$	Maximum allowed tree depth
$\alpha = 7$	Number of feature channels 3 channels for stratification splits
$N_{\text{pos}} = 1.5 \times 10^6$	Number of positive training samples
$N_{\text{neg}} = 1.5 \times 10^6$	Number of negative training samples

each atlas can be accurately mapped into target image space within a minute.

4.4 Discussion

The proposed stratified forest method is compared with the MSL classifier based landmark localisation approaches [160, 290]. In [161], RV insert points localisation results on 2D short axis images are reported, which were evaluated on the same CAP testing dataset as the one used in our experiments (*Dataset2*). The results presented in Section 4.3.1 are recomputed in 2D space by projecting the ground-truth and detected landmark points to the corresponding short axis plane. The corresponding landmark localisation results and computation time per 2D slice are shown in Table 4.6.

Similarly, the regression only forest is tested on the same image datasets, and the localisation results are observed to be less accurate than the Hough forest based localisation results. The difference can be explained with the fact that in Hough forest landmark votes are collected only from the heart surface, and this produces more consistent landmark position hypotheses. However, regression only approaches have been used in medical image analysis, particularly in [56, 91] to detect bounding boxes around the abdominal organs in CT images.

Table 4.6: Localisation accuracy of RV insertion points on 2D short axis slices. The proposed stratified forest is compared against the boosting tree classifier based landmark localisation method [161] in terms of computation time and localisation accuracy. The methods are benchmarked on the same dataset (*Dataset2*) that was used in the STACOM’12 LV Landmark Detection Challenge. The name of the datasets are abbreviated as *DS1* and *DS2* in the table.

Evaluation Dataset and Method		Mean $\pm$ Std (mm)	Median (mm)	Time (sec)
<i>DS1</i>	Stratified Forest	4.63 $\pm$ 3.81	3.57	0.11
<i>DS2</i>	MSL PBT [161]	7.90 $\pm$ 11.50	4.70	0.14
	Stratified Forest	5.55 $\pm$ 3.66	4.42	0.11

Other multi-atlas label fusion techniques [5, 268] could be chosen to segment cardiac MR images. These methods showed slight performance improvement over conventional majority voting and patch based fusion techniques when they were used to segment brain MR [5] and abdominal CT [268] images. In particular, the key-point transfer segmentation method eliminates the need for non-rigid image registration between target and atlas images, which basically reduces the dependency of segmentation algorithm on prior image alignment accuracy.

Moreover, in the experiments, a graph based regularization is tested on Hough vote images (in total 6 channels) as a post-processing step to constrain the landmark detections based on a learned inter-landmark distance model. However, we have not observed an improved performance by taking this approach. This can be explained with the joint training of all landmarks, where the regression training objective minimizes the determinant of covariance matrix in a joint manner for all the landmarks. Therefore, our approach does not require this type of post-processing as in [73].

During the inference, the proposed stratified forest does not predict a single deterministic orientation and scale value as in the case of the cascaded pose estimation methods. Testing samples can be routed in leaf nodes with different pose parameter models, which enables a probabilistic modelling of the latent variable. This can be considered as another improvement over the marginal space learning based approaches.

In the experiments, we performed additional tests by allowing the background long-range pixels

to cast Hough votes as well. It is observed that due to pose and shape variations, the background pixels introduce more dispersed landmark prediction maps which reduces the algorithm accuracy. For this reason, the predictions are only performed by the heart surface voxels in the images. Additionally, one could learn regression models conditioned on the label of different heart tissues. However, this would require multi-class segmentation.

4.5 Conclusion

In this chapter, a novel learning objective is proposed to learn more representative and patient specific decision forest classifier and regressor models. This new feature increases the landmark localisation accuracy, as the models trained with stratification splits are better able to cope with pose and size variations of the organs observed in the images. Moreover, the proposed method provides better guidance for the subsequent image analysis techniques. As shown in the experiments, state-of-the-art multi-atlas segmentation achieves better accuracy and displays robust performance when the proposed method is used as an initialization technique. Moreover, the proposed patient stratification approach is generic and modular; as such it can be used in any decision tree structure to achieve better classification and regression results. This includes applications to different modalities and other target organs. In that regard, the future work will investigate the use of stratified decision forests on 3D ultrasound images to identify viewing planes and organ locations. In the next two chapters, we will propose a super-resolution (SR) based image representation for cardiac MR images, which is experimentally shown to be a powerful image enhancement technique to boost the accuracy of subsequent image analysis and qualitative assessment.

Chapter 5

Learning to Generate High-Resolution Image Representations

This chapter is based on:

O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, Antonio de Marvao, S. Cook, D. O'Regan, and D. Rueckert, “Multi-Input Cardiac Image SuperResolution using Convolutional Neural Networks”, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Athens, Greece, 2016.

Abstract: 3D cardiac MR imaging enables accurate analysis of cardiac morphology and physiology. However, due to the requirements for long acquisition and breath-hold, the clinical routine is still dominated by multi-slice 2D imaging, which hamper the visualization of anatomy and quantitative measurements as relatively thick slices are acquired. As a solution, we propose a novel image super-resolution (SR) approach that is based on a residual convolutional neural network (CNN) model. It reconstructs high resolution 3D volumes from 2D image stacks for more accurate image analysis. The proposed model allows the use of multiple input data acquired from different viewing planes for improved performance. Experimental results on 1233

cardiac short and long-axis MR image stacks show that the CNN model outperforms state-of-the-art SR methods in terms of image quality while being computationally efficient. Also, we show that image segmentation and motion tracking benefits more from SR-CNN when it is used as an initial upscaling method than conventional interpolation methods for the subsequent analysis.

5.1 Introduction

3D magnetic resonance (MR) imaging with near isotropic resolution provides a good visualization of cardiac morphology, and enables accurate assessment of cardiovascular physiology. However, 3D MR sequences usually require long breath-hold and repetition times, which leads to scan times that are infeasible in clinical routine, and 2D multi-slice imaging is used instead. Due to limitations on signal-to-noise ratio (SNR), the acquired slices are usually thick compared to the in-plane resolution and thus negatively affect the visualization of anatomy and hamper further analysis. Attempts to improve image resolution are typically carried out either during the acquisition stage (sparse k-space filling) or retrospectively through super resolution (SR) of single/multiple image acquisitions.

5.1.1 Related work

Most of the SR methods recover the missing information through the examples observed in training images, which are used as a prior to link low and high resolution (LR-HR) image patches. Single image SR methods, based on the way they utilize training data, fall into two categories: non-parametric and parametric. The former aims to recover HR patches from LR ones via a co-occurrence prior between the target image and external training data. Atlas-based approaches such as the patch-match method [243] and non-local means based single image SR [171] methods are two examples of this category. These approaches are computationally demanding as the candidate patches have to be searched in the training dataset to find the

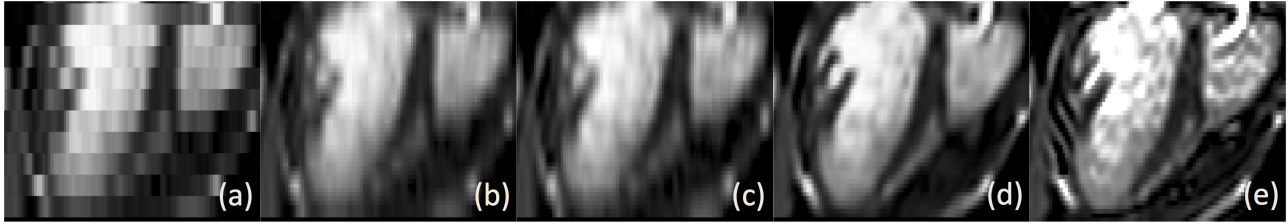


Figure 5.1: The low resolution image (a) is upscaled using linear (b) and cubic spline (c) interpolations, and the proposed method (d) which shows a high correlation with the ground-truth high resolution image (e) shown on the rightmost.

most suitable HR candidate. Instead, compact and generative models can be learned from the training data to define the mapping between LR and HR patches. Parametric generative models, such as coupled-dictionary learning based approaches, have been proposed to upscale MR brain [228] and cardiac [23] images. These methods benefit from sparsity constraint to express the link between LR and HR. Similarly, random forest based non-linear regressors have been proposed to predict HR patches from LR data and have been successfully applied on diffusion tensor images [2]. Recently, convolutional neural network (CNN) models [69, 70] have been put forward to replace the inference step as they have enough capacity to perform complex nonlinear regression tasks. Even by using a shallow network composed of a few layers, these models [70] achieved superior results over other state-of-the-art SR methods.

5.1.2 Contributions

In the work presented here, we extend the SR-CNN proposed by [69, 70] with an improved layer design and training objective function, and show its application to cardiac MR images. In particular, the proposed approach simplifies the LR-HR mapping problem through residual learning and allows training a deeper network to achieve improved performance. Additionally, the new model can be considered more data-adaptive since the initial upscaling is performed by learning a deconvolution layer instead of a fixed kernel [70]. More importantly, a multi-input image extension of the SR-CNN model is proposed and exploited to achieve a better SR image quality. By making use of multiple images acquired from different slice directions one can further improve and constrain the HR image reconstruction. Similar multi-image

SR approaches have been proposed in [188, 208] to synthesize HR cardiac images; however, these approaches did not make use of available large training datasets to learn the appearance of anatomical structures in HR. Compared to the state-of-the-art image SR approaches [70, 243], the proposed method shows improved performance in terms of peak signal-to-noise-ratio (PSNR) and structural similarity index measure (SSIM) [269]. We show that cardiac image segmentation can benefit from SR-CNN as the segmentations generated from super-resolved images are shown to be similar to the manual segmentations on HR images in terms of volume measures and surface distances. Lastly, we show that cardiac motion tracking results can be improved using SR-CNN as it visualizes the basal and apical parts of the myocardium more clearly compared to the conventional interpolation methods (see Figure 5.1).

5.2 Methodology

The SR image generation is formulated as an inverse problem that recovers the high dimensional data through the MR image acquisition model [103], which has been the starting point of approaches in [23, 188, 243]. The model links the HR volume $\mathbf{y} \in \mathbb{R}^M$ to the low dimensional observation $\mathbf{x} \in \mathbb{R}^N$ ($N \ll M$) through the application of a series of operators as: $\mathbf{x} = \mathbf{DBSM}\mathbf{y} + \boldsymbol{\eta}$ where $\mathbf{M}$ defines the spatial displacements caused due to respiratory and cardiac motion, $\mathbf{S}$ is the slice selection operator, $\mathbf{B}$ is a point-spread function (PSF) used to blur the selected slice, $\mathbf{D}$ is a decimation operator, and $\boldsymbol{\eta}$ is the Rician noise model. The solution to this inverse problem estimates a conditional distribution $p(\mathbf{y}|\mathbf{x})$ that minimizes the cost function Ψ defined by $\mathbf{y}$ and its estimate $\Phi(\mathbf{x}, \boldsymbol{\theta})$ obtained from LR input data. The estimate is obtained through a CNN parametrised by $\boldsymbol{\theta}$ that models the distribution $p(\mathbf{y}|\mathbf{x})$ via a collection of hidden variables. For the smooth ℓ_1 norm case, the loss function is defined as follows:

$$\min_{\boldsymbol{\theta}} \sum_i \Psi_{\ell_1}(\Phi(\mathbf{x}_i, \boldsymbol{\theta}) - \mathbf{y}_i) \quad (5.1)$$

$$\Psi_{\ell_1}(r) = \{0.5 r^2 \text{ if } |r| < 1, |r| - 0.5 \text{ otherwise}\}$$

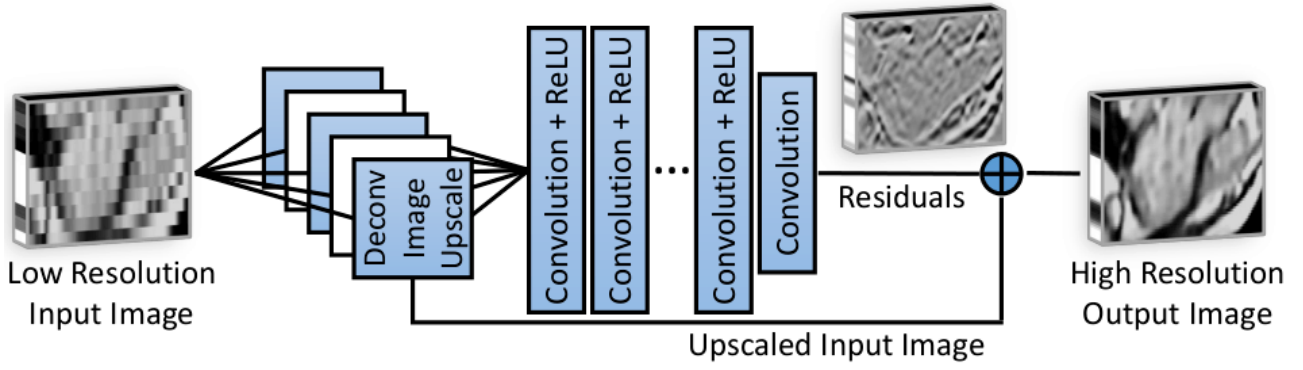


Figure 5.2: The proposed single image super resolution network model

Here $(\mathbf{x}_i, \mathbf{y}_i)$ denote the training samples. The next section describes the proposed CNN model.

5.2.1 Single Image Network

The proposed model, shown in Figure 5.2, is formed by concatenating a series of convolutional layers (Conv) and rectified linear units (ReLU) [139] to estimate the non-linear mapping Φ , as proposed in [70] to upscale natural images. The intermediate feature maps $h_j^{(n)}$ at layer n are computed through Conv kernels (hidden units) w_{kj}^n as $\max\left(0, \sum_{k=1}^K h_k^{(n-1)} * w_{kj}^n\right) = h_j^n$ where $*$ is the convolution operator. As suggested by [246], in order to obtain better non-linear estimations, the proposed architecture uses small Conv kernels (3x3x3) and a large number of Conv+ReLU layers. Such approach allows training of a deeper network. Different to the models proposed in [69, 70], we include an initial upscaling operation within the model as a deconvolution layer (Deconv) $(\mathbf{x} \uparrow U) * w_j = h_j^0$ where $\uparrow$ is a zero-padding upscaling operator and $U = M/N$ is the upscaling factor. In this way, upsampling filters can be optimized for SR applications by training the network in an end-to-end manner. This improves the image signal quality in image regions closer to the boundaries. Instead of learning to synthesize a HR image, the CNN model is trained to predict the residuals between the LR input data and HR ground-truth information. These residuals are later summed up with the linearly upscaled input image (output of Deconv layer) to reconstruct the output HR image. In this way, a simplified regression function Φ is learned where mostly high frequency signal components, such as edges

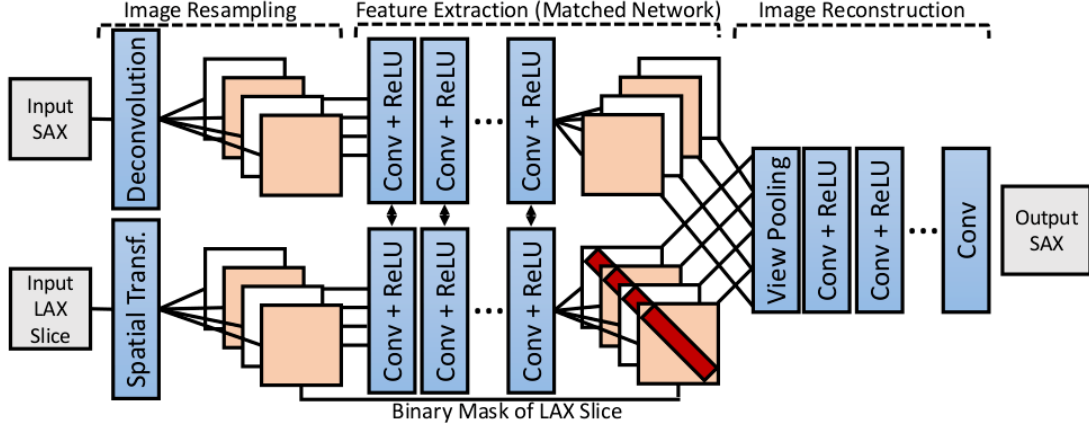


Figure 5.3: The proposed Siamese multi-image super resolution network model.

and texture, are predicted (see Fig 5.2). At training time, the correctness of reconstructed HR images is evaluated based on the $\Psi_{\ell_1}(\cdot)$ function, and the model weights are updated by back-propagating the error defined by that function. In [285] the ℓ_1 norm was shown to be a better metric than the ℓ_2 norm for image restoration and SR problems. This is attributed to the fact that the weight updates are not dominated by the large prediction errors.

5.2.2 Multi-Image Network

The single image model is extended to multi-input image SR by creating multiple input channels (MC) from given images which are resampled to the same spatial grid and visualize the same anatomy. In this way, the SR performance is enhanced by merging multiple image stacks, e.g. long-axis (LAX) and short axis (SAX) stacks, acquired from different imaging planes into a single SR volume. However, when only a few slices are acquired, a mask or distance map is required as input to the network to identify the missing information. Additionally, the number of parameters is supposed to be increased so that the model can learn to extract in image regions where the masks are defined, which increases the training time accordingly. For this reason, a Siamese network [32] is proposed as a third model (see Figure 5.3) for comparison purposes, which was used in similar problems such as shape recognition from multiple images [255]. The first stage of the network resamples the input images into a fixed HR spatial grid. In the second stage the same type of image features are extracted from each channel which

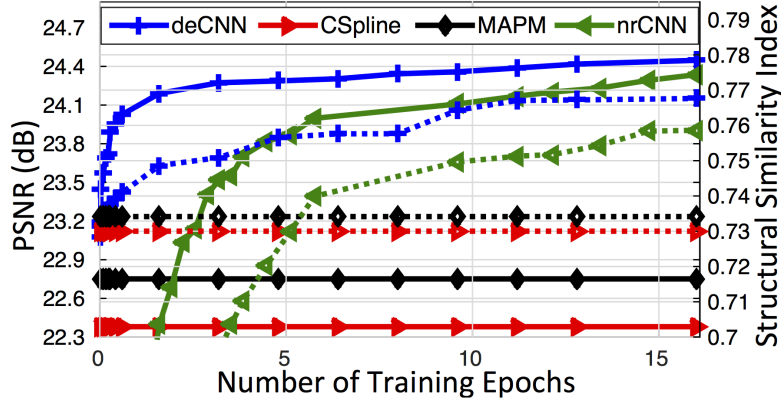


Figure 5.4: Results on the testing data, PSNR (solid) and SSIM (dashed)

are sharing the same filter weights. In the final stage, the features are pooled and passed to another Conv network to reconstruct the output HR image. The view pooling layer averages the corresponding features from all channels over the areas where the images are overlapping. The proposed models are initially pre-trained with small number of layers to better initialize the final deeper network training, which improves the network performance [69].

5.3 Experiments and Results

The models are evaluated on end-diastolic frames of cine cardiac MR images acquired from 1233 healthy adult subjects. The images are upsampled in the direction orthogonal to the SAX plane. The proposed method is compared against linear, cubic spline, and multi-atlas patchmatch (MAPM) [243] upscaling methods in four different experiments: image quality assessment for (a-b) single and multi-input cases, (c) left-ventricle (LV) segmentation, (d) LV motion tracking.

Experimental details: In the first experiment, an image dataset containing 1080 3D SAX cardiac volumes with voxel size 1.25x1.25x2.00 mm, is randomly split into two subsets and used for single-image model training (930) and testing (150). The images are intensity normalized and cropped around the heart. Synthetic LR images are generated using the acquisition model given in Section 5.2, which are resampled to a fixed resolution 1.25x1.25x10.00 mm. The PSF is set to be a Gaussian kernel with a full-width at half-maximum equal to the slice thickness [103].

Table 5.1: Quantitative comparison of different image upsampling methods.

Exp (a)	PSNR (dB)	SSIM	# Filters/Atlases
Linear	20.83±1.10	.70±.03	–
CSpline	22.38±1.13	.73±.03	–
MAPM	22.75±1.22	.73±.03	350
sh-CNN	23.67±1.18	.74±.02	64,64,32,1
CNN	24.12±1.18	.76±.02	64,64,32,16,8,4,1
de-CNN	24.45±1.20	.77±.02	64,64,32,16,8,4,1

For the LR/HR pairs, multiple acquisitions could be used as well, but an unbalanced bias would be introduced near sharp edges due to spatial misalignments. For the evaluation of multi-input models, a separate clinical dataset of 153 image pairs of LAX cardiac image slices and SAX image stacks are used, of which 10 pairs are split for evaluation. Spatial misalignment between SAX and LAX images are corrected using image registration [156]. For the single/multi image model, seven consecutive Conv layers are used after the upscaling layer. In the Siamese model, the channels are merged after the fourth Conv layer.

Image Quality Assessment: The upscaled images are compared with the ground-truth HR 3D volumes in terms of PSNR and SSIM [269]. The latter measure assesses the correlation of local structures and is less sensitive to image noise. The results in Table 5.1 show that learning the initial upscaling kernels (de-CNN) can improve ($p = .007$) the quality of generated HR image compared to convolution only network (CNN) using the same number of trainable parameters. Additionally, the performance of 7-layer network is compared against the 4-layer shallow network from [70] (sh-CNN). Addition of extra Conv layers to the 7-layer model is found to be ineffective due to increased training time and negligible performance improvement. In Fig. 5.4, we see that CNN based methods can learn better HR synthesis models even after a small number of training epochs. On the same figure, it can be seen that the model without the residual learning (nrCNN) underperforms and requires a large number of training iterations.

Multi-Input Model: In the second experiment, we show that the single image SR model can be enhanced by providing additional information from two and four chamber (2/4CH) LAX images. The results given in Table 5.2 show that by including LAX information in the model,

Table 5.2: Image quality results obtained with three different models: single-image de-CNN, Siamese, and multi-channel (MC) that uses multiple input images.

Exp (b)	de-CNN(SAX)	Siamese(SAX/4CH)	MC(SAX/4CH)	MC(SAX/2/4CH)
PSNR (dB)	24.76±0.48	25.13±0.48	25.15±0.47	25.26±0.37
SSIM	.807±.009	.814±.013	.814±.012	.818±.012
p - values	0.005	0.016	0.017	-

Table 5.3: Segmentation results for different upsampling methods, CSpline ($p = .007$) and MAPM ($p = .009$). They are compared in terms of mean and Hausdorff distances (MYO) and LV cavity volume differences (w.r.t. manual annotations).

		Linear	CSpline	MAPM	de-CNN	High Res
Exp (c)	LV Vol Diff (ml)	11.72±6.96	10.80±6.46	9.55±5.42	9.09±5.36	8.24±5.47
	Mean Dist (mm)	1.49±0.30	1.45±0.29	1.40±0.29	1.38±0.29	1.38±0.28
	Haus Dist (mm)	7.74±1.73	7.29±1.63	6.83±1.61	6.67±1.77	6.70±1.85

a modest improvement in image visual quality can be achieved. The improvement is mostly observed in image regions closer to areas where the SAX-LAX slices overlap, as can be seen in Figure 5.5 (a-d). Also, the results show that the multi-channel (MC) model performs slightly better than Siamese model as it is given more degrees-of-freedom, whereas the latter is more practical as it trains faster and requires fewer trainable parameters.

Segmentation Evaluation: As a subsequent image analysis, 18 SAX SR images are segmented using a state-of-the-art multi-atlas method [11]. The SR images generated from clinical 2D stack data with different upscaling methods are automatically segmented and those segmentations are compared with the manual annotations performed on ground-truth HR 3D images. Additionally, the HR images are segmented with the same method to show the lower error bound. The quality of segmentations are evaluated based on the LV cavity volume measure and surface-to-surface distances for myocardium (MYO). The results in Table 5.3 show that CNN upscaled images can produce segmentation results similar to the ones obtained from HR images. The main result difference between the SR methods is observed in image areas where thin and detailed boundaries are observed (e.g. apex). As can be seen in Figure 5.5 (e-h), the MAPM over-smooths areas closer to image boundaries. Inference of the proposed model is not as computationally demanding as brute-force searching (MAPM), which requires hours for a

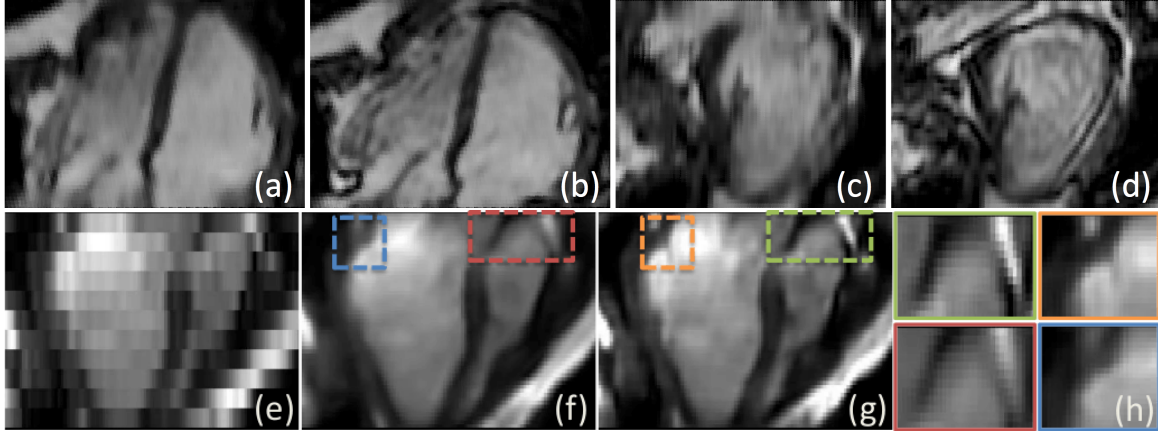


Figure 5.5: The LV is better visualized by using multi-input images (b,d) compared to single image SR (a,c). Also, the proposed method (g) performs better than MAPM [243] (f) in areas where uncommon shapes are over-smoothed by atlases.

single image, whereas SR-CNN can be executed in 6.8 s on GPU or 5.8 mins CPU on average per image. The shorter runtime makes the SR methods more applicable to subsequent analysis, as they can replace the standard interpolation methods.

Motion Tracking: The clinical applications of SR can be extended to MYO tracking as it can benefit from SR as a preprocessing stage to better highlight the ventricle boundaries. End-diastolic MYO segmentations are propagated to end-systolic (ES) phase using B-Spline FFD registrations [227]. ES meshes generated with CNN and linear upscaling methods are compared with tracking results obtained with 10 3D-SAX HR images based on Hausdorff distance. The proposed SR method produces tracking results (4.73 ± 1.03 mm) more accurate ($p = 0.01$) than the linear interpolation (5.50 ± 1.08 mm). We observe that the images upsampled with the CNN model follow the apical boundaries more accurately, which is shown in the supplementary material: www.doc.ic.ac.uk/~oo2113/M16

5.4 Discussion and Conclusion

The results show that the proposed SR approach outperforms conventional upscaling methods both in terms of image quality metrics and subsequent image analysis accuracy. Also, it is

computationally efficient and can be applied to image analysis tasks such as segmentation and tracking. The experiments show that these applications can benefit from SR images since 2D stack image analysis with SR-CNN can achieve similar quantitative results as the analysis on isotropic volumes without requiring long acquisition time. We also show that the proposed model can be easily extended to multiple image input scenarios to obtain better SR results. SR-CNN's applicability is not only limited to cardiac images but to other anatomical structures as well. In the proposed approach, inter-slice and stack spatial misalignments due to motion are handled using a registration method. However, we observe that large slice misplacements can degrade SR accuracy. Future research will focus on that aspect of the problem.

Chapter 6

Learning Anatomical Shape Priors through Neural Networks

This chapter is based on:

O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. Cook, A. Marvao, T. Dawes, D. O'Regan, B. Kainz, B. Glocker, and D. Rueckert, “Anatomically Constrained Neural Networks (ACNN): Application to Cardiac Image Enhancement and Segmentation”, IEEE Transactions on Medical Imaging (TMI), In print, Aug 2017.

Abstract: Incorporation of prior knowledge about organ shape and location is key to improve performance of image analysis approaches. In particular, priors can be useful in cases where images are corrupted and contain artefacts due to limitations in image acquisition. The highly constrained nature of anatomical objects can be well captured with learning based techniques. However, in most recent and promising techniques such as CNN based segmentation it is not obvious how to incorporate such prior knowledge. State-of-the-art methods operate as pixel-wise classifiers where the training objectives do not incorporate the structure and inter-dependencies of the output. To overcome this limitation, we propose a generic training strategy that incor-

porates anatomical prior knowledge into CNNs through a new regularisation model, which is trained end-to-end. The new framework encourages models to follow the global anatomical properties of the underlying anatomy (*e.g.* shape, label structure) via learned non-linear representations of the shape. We show that the proposed approach can be easily adapted to different analysis tasks (*e.g.* image enhancement, segmentation) and improve the prediction accuracy of the state-of-the-art models. The applicability of our approach is shown on multi-modal cardiac datasets and public benchmarks. Additionally, we demonstrate how the learned deep models of 3D shapes can be interpreted and used as biomarkers for classification of cardiac pathologies

6.1 Introduction

Image segmentation techniques aim to partition an image into meaningful parts which are used for further analysis. The segmentation process is typically driven by both the underlying data and a prior on the solution space, where the latter is useful in cases where the images are corrupted or contain artefacts due to limitations in the image acquisition. For example, bias fields, shadowing, signal drop-out, respiratory motion, and low-resolution acquisitions are the few common limitations in ultrasound (US) and magnetic resonance (MR) imaging.

Incorporating prior knowledge into image segmentation algorithms has proven useful in order to obtain more accurate and plausible results as summarised in the recent survey [187]. Prior information can take many forms: boundaries and edge polarity [44]; shape models [52, 62]; topology specification; distance prior between regions; atlas models [11], which were commonly used as a regularisation term in energy optimisation based traditional segmentation methods (*e.g.* region growing). In particular, atlas priors are well suited for medical imaging applications since they enforce both location and shape priors through a set of annotated anatomical atlases. Similarly, auto-context models [266] have made use of label and image priors in segmentation, which require a cascade of models.

In the context of neural networks (NNs), early work on shape analysis has focused on learning

generative models through deep Boltzmann Machines (DBMs), namely ShapeBM [77] that uses a form of DBM with sparse pixel connectivity. Follow-up work in [43, 76] has demonstrated the application of DBMs to binary segmentation problems in natural images containing vehicles and other types of objects. However, fully connected DBM for images require a large number of parameters and consequently model training may become intractable depending on the size of images. For this reason, convolutional deep belief nets [273] were recently proposed for encoding shape prior information. Besides variational models, cascaded convolutional architectures [146, 217] have been shown to discover priors on shape and structure in label space without any a priori specification. However, this comes at the cost of increased model complexity and computational needs.

In the context of medical imaging and neural networks, anatomical priors have not been studied in much depth, particularly in the current state-of-the-art segmentation techniques [45, 123, 218, 225]. Recent work has shown simple use cases of priors through adjacency [19] and boundary [44] conditions. Inclusion of priors in medical imaging could potentially have much more impact compared to their use in natural image analysis since anatomical objects in medical images are naturally more constrained in terms of their shape and location.

As explained in a recent NN survey paper [150], the majority of the classification and regression models utilise a pixel-level loss function (*e.g.* cross-entropy or mean square error) which does not fully take into account the underlying semantic information and dependencies in the output space (*e.g.* class labels). In this chapter, we present a novel and generic way to incorporate global shape/label information into NNs. The proposed approach, namely anatomically constrained neural networks (ACNN), is mainly motivated by the early work on shape priors and image segmentation, in particular PCA based statistical [52] and active shape models [62]. Our framework learns a non-linear compact representation of the underlying anatomy through a stacked convolutional autoencoder [175] and enforces network predictions to follow the learnt statistical shape/label distributions. In other words, it favours predictions that lie on the extracted low dimensional data manifold. More importantly, our approach is independent of the

particular NN architecture or application; it can be combined with any of the state-of-the-art segmentation or super-resolution (SR) NN models and potentially improve its prediction accuracy and robustness without introducing any memory or computational complexity at inference time. Lastly, ACNN models, trained with the proposed prior term which acts as a regulariser, remove the need for post-processing steps such as conditional random fields [135] which are often based on heuristics parameter tuning. In ACNN, the regularisation is part of the end-to-end learning which can be a great advantage.

The proposed global training objective in SR corresponds to a prior on the space of feasible high-resolution (HR) solutions, which is experimentally shown to be useful since SR is an ill-posed problem. Similar modifications of the objective function during training have been introduced to enhance the quality of natural images, such as perceptual [120] and adversarial [141] loss terms, which were used to synthesise more realistic images in terms of texture and object boundaries. In the context of medical imaging, our priors enforce the synthesised HR images to be anatomically meaningful while minimising a traditional image reconstruction loss function.

6.1.1 Clinical Motivation

Cardiac imaging has an important role in diagnosis, pre-operative planning, and post-operative management of patients with heart disease. Imaging modalities such as US and cardiac MR (CMR) are widely used to provide detailed assessment of cardiac function and morphology. Each modality is suitable for particular clinical use cases; for instance, 2D-US is still the first line of choice due to its low cost and wide availability, whereas, CMR is a more comprehensive modality with excellent contrast for both anatomical and functional evaluation of the heart [124]. Similarly, 3D-US is recommended over the use of 2D-US since it has been demonstrated to provide more accurate and reproducible volumetric measurements [137].

Some of the standard clinical acquisition protocols in 3D-US and CMR still have limitations

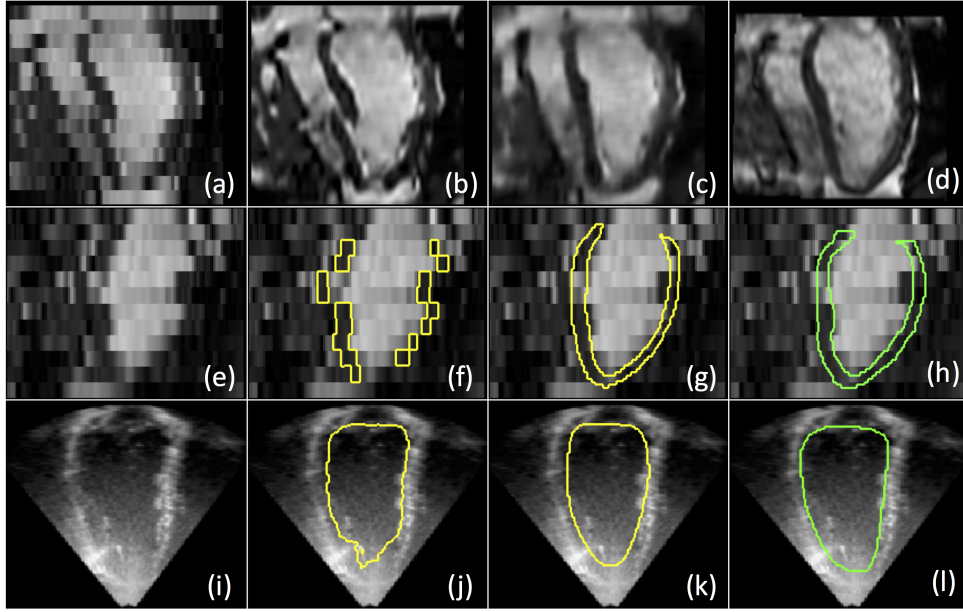


Figure 6.1: Results for cardiac MR super-resolution (SR) (top), MR segmentation (middle), and ultrasound (US) segmentation (bottom). From left to right, we show the input image, a state-of-the-art competing method, the proposed result, and the ground-truth. (a) Stack of 2D MR images with respiratory motion artefacts, (b) SR based on CNNs [191], (c) the proposed ACNN-SR, (d) ground-truth high-resolution (HR) image, (e) low resolution MR image, (f) 2D segmentation resulting in blocky contours [264], (g) 3D sub-pixel segmentation from stack of 2D MR images using ACNN, (h) manual segmentation from HR image, (i) input 3D-US image, (j) FCN based segmentation [45], (k) ACNN, and (l) manual segmentation.

in visualising the underlying anatomy due to imaging artefacts (*e.g.* cardiac motion, low slice resolution, lack of slice coverage [206]) or operator-dependent errors (*e.g.* shadows, signal drop-outs). In the clinical routine, these challenges are usually tackled through multiple acquisitions of the same anatomy and repeated patient breath-holds leading to long examination times. Similar problems have been reported in large cohort studies such as the UK Biobank [206], which leads to inaccurate quantitative measurements or even the discarding of acquired images. As can be seen in Figure 6.1, the existing state-of-the-art convolutional neural network (CNN) approaches for segmentation [45, 264] and image enhancement [191] tasks perform poorly when the input data is not self-consistent for the analysis. For this reason, incorporation of prior knowledge into cardiac image analysis could provide more accurate and reliable assessment of the anatomy, which is shown in the third column of the same figure. Most importantly, the proposed ACNN model allows us to perform HR analysis via sub-pixel feature maps generated

from low resolution (LR) input data even in the presence of motion artefacts. Using the proposed approach we can perform full 3D segmentation without explicit motion correction and do not have to rely on LR slice-by-slice 2D segmentation.

We demonstrate the applicability of the proposed approach for cine stacks of 2D MR and 3D-US datasets composed of 1200 and 45 cardiac image sequences respectively. We show that the proposed segmentation and SR models become more robust against imaging artefacts mentioned earlier which is underlined by our state-of-the-art results on the MICCAI'14 CETUS public benchmark [20]. We also demonstrate that the lower dimensional representations learnt by the proposed ACNN can be useful for classification of pathologies such as dilated and hypertrophic cardiomyopathy, and it does not require point-wise correspondence search between subjects as in [239]. For the evaluation, the MICCAI'17 AC/DC classification benchmark was used. In that regard, the proposed method is not only useful for image enhancement and segmentation but also for the study of anatomical shape variations in population studies and their associations with cardiac related pathologies.

6.1.2 Contributions

In this study, we propose a generic and novel technique to incorporate priors on shape and label structure into NNs for medical image analysis tasks. In this way, we can constrain the NN training process and guide the NN to make anatomically more meaningful predictions, in particular in cases where the input image data is not informative or consistent enough (*e.g.* missing object boundaries). More importantly, to the best of our knowledge, this is one of the earliest studies demonstrating the use of convolutional autoencoder networks to learn anatomical shape variations from medical images.

The proposed ACNN model is evaluated on multi-modal cardiac datasets from MR and US. Our evaluation shows: (I) A sub-pixel cardiac MR image segmentation approach that, in contrast to previous CNN approaches [6, 264], is robust against slice misalignment and coverage problems;

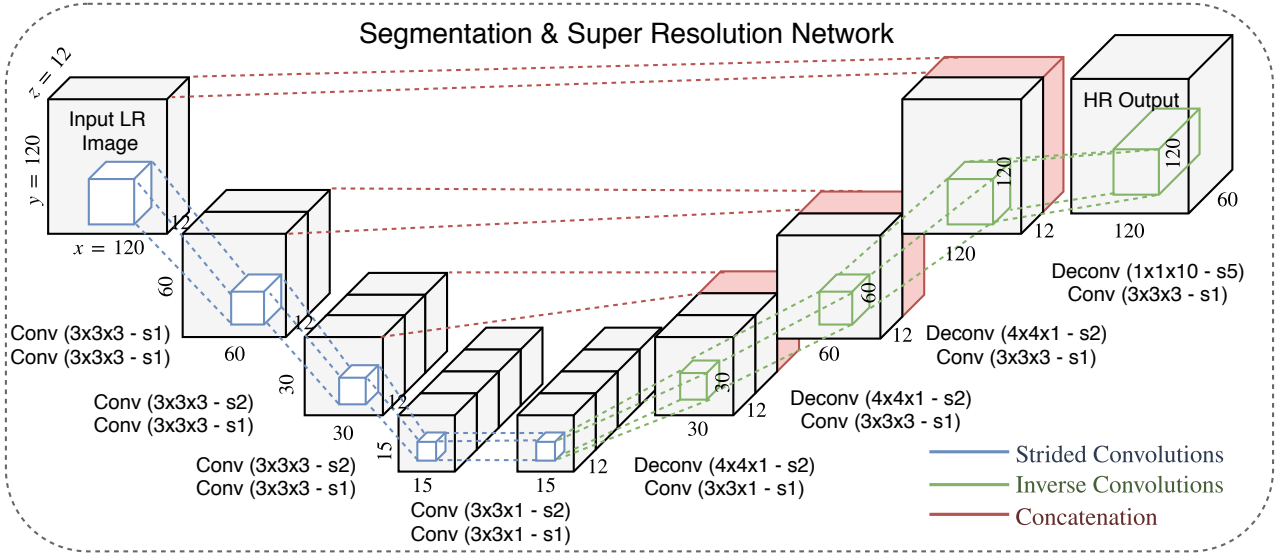


Figure 6.2: Block diagram of the baseline segmentation (Seg) and super-resolution (SR) models which are combined with the proposed T-L regularisation block (shown in Figure 6.3) to build the ACNN-Seg/SR frameworks. In SR, the illustrated model extracts SR features in low-resolution (LR) space, which increases computational efficiency. In segmentation, the model achieves sub-pixel accuracy for given LR input image. The skip connections between the layers are shown in red.

(II) An implicit statistical parametrisation of the left ventricular shape via NNs for pathology classification; (III) An image SR technique that extends previous work [191] and that is robust against slice misalignments; our approach is computationally more efficient than the state-of-the-art SR-CNN model [191] as the feature extraction is performed in the low-dimensional image space. (IV) Last, we demonstrate state-of-the-art 3D-US cardiac segmentation results on the CETUS'14 Benchmark.

6.2 Methodology

In the next section, we briefly summarise the state-of-the-art methodology for image segmentation (SEG) and super-resolution (SR), which is based on convolutional neural networks (CNNs). We then present a novel methodology that extends these CNN models with a global training objective to constrain the output space by imposing anatomical shape priors. For this, we propose a new regularisation network that is based on the T-L architecture which was used in

computer graphics [94] to 3D render objects from natural images.

6.2.1 Medical Image Segmentation with CNN Models

Let $\mathbf{y}_s = \{y_i\}_{i \in \mathcal{S}}$ be an image of class labels representing different tissue types with $y_i \in \mathcal{L} = \{1, 2, \dots, C\}$. Furthermore let $\mathbf{x} = \{x_i \in \mathbb{R}, i \in \mathcal{S}\}$ be the observed intensity image. The aim of image segmentation is to estimate $\mathbf{y}_s$ having observed $\mathbf{x}$. In CNN based segmentation models [123, 154, 225], this task is performed by learning a discriminative function that models the underlying conditional probability distribution $P(\mathbf{y}_s|\mathbf{x})$.

The estimation of class densities $P(\mathbf{y}_s|\mathbf{x})$ consists in assigning to each x_i the probability of belonging to each of the C classes, yielding C sets of class feature maps h_c that are extracted through learnt non-linear functions. The final decision for class labels is then made by applying softmax to the extracted class feature maps, in the case of cross-entropy $L_x = -\sum_{c=1}^C \sum_{i \in \mathcal{S}} \log \left(\frac{e^{h_{(c,i)}}}{\sum_j e^{h_{(j,i)}}} \right)$ these feature maps correspond to log likelihood values.

As in the U-Net [225] and DeepMedic [123] models, we learn the mapping between intensities and labels $\phi(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{L}$ by optimising the average cross-entropy loss of each class $L_x = \sum_{c=1}^C L_{(x,c)}$ using stochastic gradient descent. As shown in Figure 6.2, the mapping function ϕ is computed by passing the input image through a series of convolution layers and rectified linear units across different image scales to enlarge the model's receptive field. The presented model is composed of two parts: feature extraction (analysis) similar to a VGG-Net [246] and reconstruction (synthesis) as in the case of a 3D U-Net [225]. However, in contrast to existing approaches, we aim for sub-pixel segmentation accuracy by training up-sampling layers with high-resolution ground-truth maps. This enables 3D analysis of the underlying anatomy in case of thick slice 2D image stack acquisitions such as cine cardiac MR imaging. In this way, it is possible to perform analysis on the high-resolution image grid without any preceding upsampling operation with a SR model [191].

Similar segmentation frameworks (cf. [150]) have been studied in medical imaging. However,

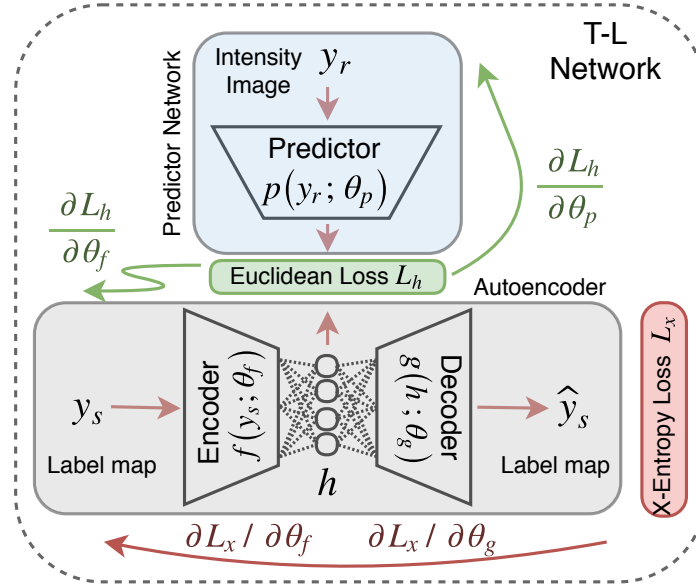


Figure 6.3: Block diagram of the stacked convolutional autoencoder (AE) network (in grey), which is trained with segmentation labels. The AE model is coupled with a predictor network (in blue) to obtain a compact non-linear representation that can be extracted from both intensity and segmentation images. The whole model is named as T-L network.

in most of the existing methods, the models are supervised purely through a local loss function at pixel level (*e.g.* cross-entropy, Dice) without exploiting the global dependencies and structure in the output space. For this reason, the global description of predictions is usually not adhering to shape, label, or atlas priors. In contrast to this we propose a model that can incorporate the aforementioned priors in segmentation models. The proposed framework relies on autoencoder and T-L network models to obtain a non-linear compact representation of the underlying anatomy, which are used as priors in segmentation.

6.2.2 Convolutional Autoencoder Model and ACNN-Seg

An autoencoder (AE) [267] is a neural network that aims to learn an intermediate representation from which the original input can be reconstructed. Internally, it has a hidden layer $\mathbf{h}$ whose activations represent the input image, often referred as *codes*. To avoid the AE to directly copy its output, the AE are often designed to be undercomplete so that the size of the code is less than the input dimension as shown in Figure 6.3. Learning an AE forces the network

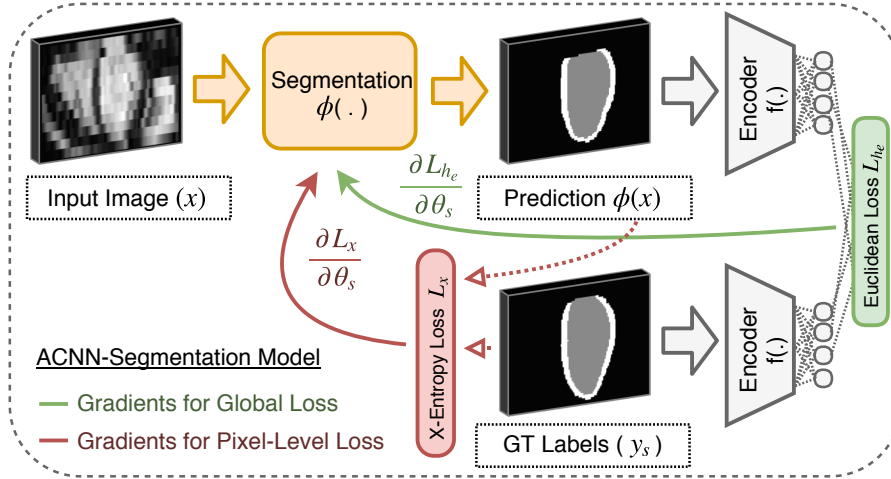


Figure 6.4: Training scheme of the proposed anatomically constrained convolutional neural network (ACNN) for image segmentation task. The encoder part of the proposed T-L network is used as a regularisation model to enforce the model predictions to follow the distribution of the learned low dimensional representations or priors.

to capture the most salient features of the training data. The learning procedure minimises a loss function $L_x(\mathbf{y}_s, g(f(\mathbf{y}_s)))$, where L_x is penalising $g(f(\mathbf{y}_s))$ being dissimilar from $\mathbf{y}_s$. The functions g and f are defined as the decoder and encoder components of the AE.

In the proposed method, the AE is integrated into the standard segmentation network, described in Sec. 6.2.1, as a regularisation model to constrain class label predictions $\mathbf{y}$ towards anatomically meaningful and accurate outputs. The cross-entropy loss function operates on individual pixel level class predictions, which does not guarantee global consistency and plausible anatomical shapes even though the segmentation network has a receptive field larger than the size of structures to be segmented. This is due to the fact that back-propagated gradients are parametrised only by pixel-wise individual probability divergence terms and thus provide little global context.

To overcome this limitation, class prediction label maps are passed through the AE to obtain a lower dimensional (*e.g.* 64 dimensions) parametrisation of the segmentation and its underlying structure [240]. By performing AE-based non-linear lower dimensional projections on both predictions and ground-truth labels, as shown in Figure 6.4, we can build our ACNN-Seg training objective function through a linear combination of cross-entropy (L_x), shape regularisation loss

(L_{he}) , and weight decay terms as follows:

$$\begin{aligned} L_{he} = & \|f(\phi(\mathbf{x}); \boldsymbol{\theta}_f) - f(\mathbf{y}; \boldsymbol{\theta}_f)\|_2^2 \\ \min_{\boldsymbol{\theta}_s} & \left(L_x(\phi(\mathbf{x}; \boldsymbol{\theta}_s), \mathbf{y}) + \lambda_1 \cdot L_{he} + \frac{\lambda_2}{2} \|\mathbf{w}\|_2^2 \right) \end{aligned} \quad (6.1)$$

Here $\mathbf{w}$ corresponds to weights of the convolution filters, and $\boldsymbol{\theta}_s$ denotes all trainable parameters of the segmentation model and only these parameters are updated during training. The coupling parameters λ_1 and λ_2 determine the weights of shape regularisation loss and weight decay terms used in the training. In this equation, the second term L_{he} ensures that the generated segmentations are in a similar low dimensional space (*e.g.* shape manifold) as the ground-truth labels. In addition to imposing shape regularisation, this parametrisation encourages label consistency in model predictions, and reduces false-positive detections as they can influence the predicted codes in the hidden layer. The third term corresponds to weight decay to limit the number of free parameters in the model to avoid over-fitting. The proposed AE model is composed of convolutional layers and a fully connected layer in the middle as shown in Figure 6.3, which is similar to the stacked convolutional autoencoder model proposed in [175]. The AE model details (*e.g.* layer configuration, parameter choices) are provided in the supplementary material.

6.2.3 Medical Image Super-Resolution (SR) with CNNs

Super-resolution (SR) image generation is an inverse problem where the goal is to recover spatial frequency information that is outside the spatial bandwidth of the low resolution (LR) observation $\mathbf{x} \in \mathbb{R}^N$ to predict a high resolution (HR) image $\mathbf{y}_r \in \mathbb{R}^M$ ($N \ll M$), as illustrated in the top row of Figure 6.1. Since the high frequency components are missing in the observation space, usually training examples are used to predict the most likely $P(\mathbf{y}_r|\mathbf{x})$ HR output. Image SR is an ill-posed problem as there are an infinite number of solutions for a given input sample but only a few would be anatomically meaningful and accurate. As for the case of image

segmentation, learnt shape representations can be used to regularise image SR, constraining the model to make only anatomically meaningful predictions.

Similar to the SR framework described in [191], our proposed SR model learns a mapping function $\Phi : \mathcal{X} \rightarrow \mathcal{Y}$ to estimate a high-resolution image $\hat{\mathbf{y}}_r = \Phi(\mathbf{x}; \boldsymbol{\theta}_r)$ where $\boldsymbol{\theta}_r$ denotes the model parameters such as convolution kernels and batch-normalisation statistics. The parameters are optimised by minimising the smooth ℓ_1 loss, also known as Huber loss, between the ground-truth high resolution image and the corresponding prediction. The smooth ℓ_1 norm is defined as $\Psi_{\ell_1}(k) = \{0.5 k^2 \text{ if } |k| < 1, |k| - 0.5 \text{ otherwise}\}$ and the SR training objective becomes $\min_{\boldsymbol{\theta}_r} \sum_{i \in \mathcal{S}} \Psi_{\ell_1}(\Phi(\mathbf{x}_i; \boldsymbol{\theta}_r) - \mathbf{y}_i)$

In the proposed SR framework, we used the same model as shown in Figure 6.2. It provides two main advantages over the state-of-the-art medical image SR model proposed in [191]: (I) the network generates image features in the LR image grid rather than early upsampling of the features, which reduces memory and computation requirements significantly. As highlighted in [242], early upsampling introduces redundant computations in the HR space since no additional information is added into the model by performing transposed convolutions [280] at an early stage. (II) The second advantage is the use of a larger receptive field to learn the underlying anatomy, which was not the case in earlier SR methods used in medical imaging [191] and natural image analysis [72, 242] because these models usually operate on local patch level. Capturing large context indeed helps our model to better understand the underlying anatomy and this enables us to enforce global shape constraints. This is achieved by generating SR feature-maps in multiple scales using multi strides in the in-plane direction.

Similar to the ACNN-Seg model, it is possible to regularise SR models to synthesise anatomically more meaningful HR images. To achieve this goal, we extend the standard AE model to the T-L model which enables us to obtain shape representation codes directly from the intensity space. The idea is motivated by the recent work [94] on 3D shape analysis in natural images. In the next section we will explain the training strategy and the use of the T-L model as a regulariser.

6.2.4 T-L Network Model and SR-ACNN

Shape encoding AE models operate only on the segmentation masks and this limits its application to SR problem where the model output is an intensity image. To circumvent this problem, we extend the standard denoising AE to the T-L regularisation model by combining the AE with a predictor network (Figure 6.3) $p(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{H}$. The predictor can map an input image into a low dimensional non-parametric representation of the underlying anatomy (*e.g.* shape and class label information), which is learnt by the AE. In other words, it enables us to learn a hidden representation space that can be reached by non-linear mappings from both image label space $\mathcal{Y}$ and image intensity space $\mathcal{X}$. In this way, SR models can be regularised as well with respect to learnt anatomical priors.

This network architecture is useful in image analysis applications for two main reasons: (I) It enables us to build a regularisation network that could be used in applications different than image segmentation such as image SR. We propose to use this new regularisation network at training time of SR to enforce the models to learn global information about the images besides the standard pixel-wise (ℓ_1 distance) image reconstruction loss. In this way, the regressor SR model is guided by the additional segmentation information, and it becomes robust against imaging artefacts and missing information. (II) The second important feature of the T-L model is the generalisation of the learnt representations. Joint training of the AE and predictor enables us to learn representations that could be extracted from both intensity and label space. The learnt codes will encode the variations that could be interpreted from both manual annotations and intensity images. Since a perfect mapping between the intensity and label spaces is practically not achievable, the T-L learnt codes are expected to be more representative due to the inclusion of additional information.

The T-L model is trained in two stages: In the first stage, the AE is trained separately with ground-truth segmentation masks and cross-entropy loss L_x . Later, the predictor model is trained to match the learnt latent space $\mathbf{h}$ by minimising the Euclidean distance L_h between the codes predicted by the AE and predictor as shown in Figure 6.3. Once the loss functions for

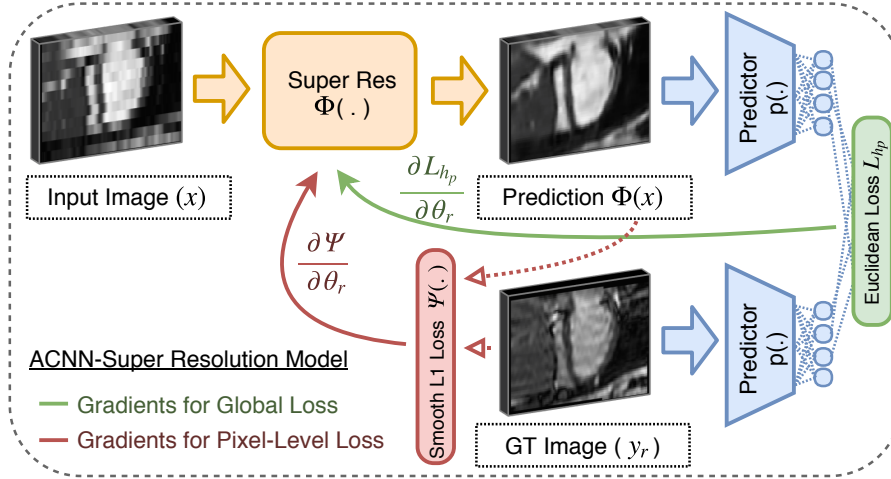


Figure 6.5: Training scheme of the proposed anatomically constrained convolutional neural network (ACNN) for image super-resolution task. The predictor part of the proposed T-L network is used as a regularisation model to enforce the model predictions to follow the distribution of the learned low dimensional representations or priors.

both the AE and the predictor converge, the two models are trained jointly in the second stage. The encoder f is updated using two separate back-propagated gradients $(\frac{\partial L_x}{\partial \theta_f}, \frac{\partial L_h}{\partial \theta_f})$ and the two loss functions are scaled to match their range. The first gradient encourages the encoder to generate codes that could be easily extracted by the predictor while the second gradient making sure that a good segmentation-reconstruction can be obtained at the output of the decoder. Training details are further discussed in Section 6.3.2. It is important to note that the T-L regulariser model is used only at training time but not during inference; in other words, the fully convolutional (FCN) segmentation and super-resolution models can still be used for applications using different image sizes. In this chapter, the proposed SR model is referred to as ACNN-SR and its training scheme is shown in Figure 6.5.

$$L_{hp} = \|p(\Phi(x); \theta_p) - p(y_r; \theta_p)\|_2^2$$

$$\min_{\theta_r} \left(\Psi_{\ell_1}(\Phi(x; \theta_r) - y_r) + \lambda_1 \cdot L_{hp} + \frac{\lambda_2}{2} \|\mathbf{w}\|_2^2 \right) \quad (6.2)$$

The training objective shown above is composed of weight decay, pixel-wise and global loss terms. Here λ_1 and λ_2 determine the weight of shape priors and weight decay terms while the smooth ℓ_1 norm loss function Ψ quantifies the reconstruction error. The global loss L_{hp}

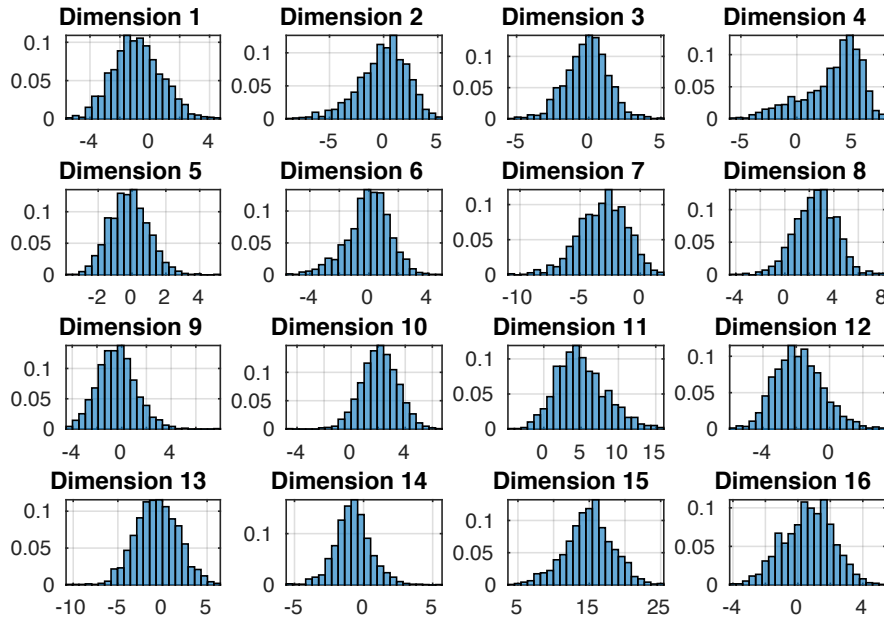


Figure 6.6: Histogram of the learnt low-dimensional latent representations (randomly selected 16 components are shown). The codes in general follow a smooth and normal distribution which is important for the training of ACNN models.)

is defined as the Euclidean distance between the codes generated from the synthesised and ground-truth HR images. The T-L model is used only in the network training phase as a regularisation term, similar to VGG features [246] that were used for representing a perceptual loss function [120]. However, we are not interested in expanding the output space to a larger feature-map space, but instead obtain a compact representation of the underlying anatomy.

6.2.5 Learnt Hidden Representations

The learnt low dimensional representation $\mathbf{h}$ is used to constrain NN models. Low dimensional encoding enables us to train models with global characteristics but also yields better generalisation power for the underlying anatomy as shown in earlier work [263]. However, since we update our segmentation and SR model parameters with the gradients back-propagated from the global loss layer using the Euclidean distance of these representations, it is essential to analyse the distribution of the extracted codes. In Figure 6.6, due to space limitations, we show the histogram of 16 randomly chosen codes (out of 64) of a T-L model trained with cardiac MR

segmentations. Note that each histogram is constructed using the corresponding code for every sample in the full dataset. It is observed that the learnt latent representations in general follow a normal distribution and they are not separated in multi-clusters (*e.g.* mixture of Gaussians). A smooth distribution of the codes ensures better supervision for the main NN model (SR, Seg) since the global gradients are back-propagated by computing the Euclidean distance between the obtained distributions.

This observation can be explained by the fact that the proposed T-L network is trained with small Gaussian input noise as in the case of denoising autoencoders. In [1], Alain and Bengio showed that the denoising reconstruction error is equivalent to contractive penalty, which forces the feature extraction (encoder) function f resist perturbations of the input and contracts these input samples to similar low dimensional codes. The penalty is defined as $\Omega(\mathbf{h}) = \lambda \left\| \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \right\|_F^2$, where F denotes the Frobenius norm (sum of squared elements), and $\mathbf{h} = f(\mathbf{x})$ represents the codes. The given penalty function promotes the network to learn the underlying low-dimensional data manifold and capture its local smooth structure. In addition to the smoothness of the latent distributions, the extracted codes are expected to be correlated since the decoder merges some of the codes along the three spatial dimensions to construct input feature maps for the transposed convolutions, but this characteristic is not a limitation in our study.

6.3 Applications and Experiments

In this section, we present three different applications of the proposed ACNN model: 3D-US and cardiac MR image segmentation, as well as cardiac MR image SR. The experiments focus on demonstrating the importance of shape and label priors for image analysis. Additionally, we analyse the salient information stored in the learnt hidden representations and correlate them with clinical indices, showing their potential use as biomarkers for pathology classification. The next subsection describes the clinical datasets used in our experiments.

6.3.1 Clinical Datasets

UK Digital Heart Project Dataset

This dataset¹ is composed of 1200 pairs of cine 2D stack short-axis (SAX) and cine 3D high resolution (HR) cardiac MR images. Each image pair is acquired from a healthy subject using a standard imaging protocol [10, 174]. In more detail, the 2D stacks are acquired in different breath-holds and therefore may contain motion artefacts. Similarly, 3D imaging is not always feasible in the standard clinical setting due to the requirements for long image acquisition. The voxel resolution of the images are fixed to 1.25x1.25x10.00 mm and 1.25x1.25x2.00 mm for 2D stack low resolution (LR) and HR images respectively. Dense segmentation annotations for HR images are obtained by manually correcting initial segmentations generated with a semi-automatic multi-atlas segmentation method [11], and all the annotations are performed on the HR images to minimise errors introduced due to LR in through plane direction. Since the ground-truth information is obtained from the HR motion-free images, the experimental results are expected to reflect the performance of the method with respect to an appropriate reference. The annotations consist of pixel-wise labelling of endocardium and myocardium classes. Additionally, the residual spatial misalignment between the 2D LR stacks and HR volumes is corrected using a rigid transformation estimated by an intensity based image registration algorithm.

CETUS'14 Challenge Dataset

CETUS'14 segmentation challenge [20] is a publicly available platform² to benchmark cardiac 3D ultrasound (US) left-ventricle (LV) segmentation methods. The challenge dataset is composed of 3D+time US image sequences acquired from 15 healthy subjects and 30 patients diagnosed with myocardial infarction or dilated cardiomyopathy. The images were acquired

¹<https://digital-heart.org/>

²<https://www.creatis.insa-lyon.fr/Challenge/CETUS/index.html>

from apical windows and LV chamber was the main focus of analysis. Resolution of the images was fixed to 1 mm isotropic voxel size through linear interpolation. The associated manual contours of the LV boundary were drawn by three different expert cardiologists, and the annotations were performed only on the frames corresponding to end-diastole (ED) and end-systole (ES) phases. Method evaluation is performed in a blinded fashion on the testing set (30 out of 45) using the MIDAS web platform.

ACDC MICCAI'17 Challenge Dataset

The aim of the ACDC'17 challenge³ is to compare the performance of automatic methods for the classification of MR image examinations in terms of healthy and pathological cases: infarction, dilated cardiomyopathy, and hypertrophic cardiomyopathy. The publicly available dataset consists of 20 (per class) cine stacks of 2D MR image sequences which are annotated at ED and ES phases by a clinical expert. In the experiments, latent representations (*codes*) extracted with the proposed T-L network are used to classify these images.

6.3.2 Training Details of the Proposed Model

In this section, we discuss the details of data augmentation used in training, and also the optimisation scheme of the T-L model training. To improve the model's generalisation capability, the input training samples are artificially augmented using affine transformations, which is used in both the segmentation and T-L models. For the SR models, on the other hand, respiratory motion artefacts between the adjacent slices are simulated via in-plane rigid transformations that are defined for each slice independently. The corresponding ground-truth HR images are not spatially transformed; in this way, the models learn to output anatomically correct results when the input slices are motion corrupted. Additionally, additive Gaussian noise is applied to input intensity images to make the segmentation and super-resolution models more robust against

³<https://www.creatis.insa-lyon.fr/Challenge/acdc/>

image noise. For the AE, the tissue class labels are randomly swapped with the probability of 0.1 to encourage the model to map slightly different segmentation masks to neighbouring points in the lower dimensional latent space. It ensures the smoothness of the learnt low-dimensional manifold space as explained in Section 6.2.5.

In the joint training of the T-L network, parameters of the encoder model (f) are updated by the gradients originating from both the cross-entropy loss (L_x) and Euclidean distance terms (L_h). Instead of applying these two gradient descent updates sequentially in an iterative fashion, we perform a joint update training scheme and experimentally observed better convergence.

6.3.3 Cardiac Cine-MR Image Segmentation

In this experiment, NN models are used to segment cardiac cine MR images in the dataset described in Sec. 6.3.1. As an input to the models, only the 2D stack LR images are used, which is a commonly used acquisition protocol for cardiac imaging, and the segmentation is performed only on the ED phase of the sequences. The corresponding ground-truth label maps, however, are projected from the HR image space, which are annotated in the HR image grid. The dataset (1200 LR images & HR labels) is randomly partitioned into three subsets: training (900), validation (100), and testing (200). All the images are linearly intensity normalised and cropped based on the automatically detected six anatomical landmark locations [190]. Other segmentation approaches employed simpler techniques (*e.g.* Hough Transform) to localise the heart in MR images.

The proposed ACNN-Seg method is compared against: the current state-of-the-art cine MR 2D slice by slice segmentation method (2D-FCN) [264], 3D-UNet model [49], cascaded 3D-UNet and convolutional AE model (AE-Seg) [217], sub-pixel 3D-CNN segmentation model (3D-Seg) proposed in Sec. 6.2.1, and the same model trained with various types of motion augmentation (3D-Seg-MAug). As the models have a different layout, the number of trainable parameters (pars) used in each model is kept fixed to avoid any bias. For the cascaded AE-Seg model,

Table 6.1: Stacks of 2D cardiac MR images (200) are segmented into LV endocardium and myocardium, and the segmentation accuracy is evaluated in terms of Dice metric and surface to surface distances. The ground-truth labels are obtained from high resolution 3D images acquired from same subjects, which do not contain motion and blocky artefacts. The proposed approach (ACNN-Seg) is compared against state-of-the-art slice by slice segmentation (2D-FCN [264]) method, 3D-UNet model [49], cascaded 3D-UNet and convolutional AE model (AE-Seg) [217], proposed sub-pixel segmentation model (3D-Seg) and the same model with motion augmentation used in training (3D-Seg-MAug).

	Method	Mean Dist. (mm)	Hausdorff Dist. (mm)	Dice Score (%)	# Trainable Parameters
<i>Endo-cardium</i>	2D-FCN [264]	2.07±0.61	11.37±7.15	.908±.021	1.39×10^6
	3D-Seg	1.77±0.84	10.28±8.25	.923±.019	1.60×10^6
	3D-UNet [49]	1.66±0.74	9.94±9.22	.923±.019	1.64×10^6
	AE-Seg [217]	1.75±0.58	8.42±3.64	.926±.019	1.68×10^6
	3D-Seg-MAug	1.59±0.74	8.52±8.13	.928±.019	1.60×10^6
	AE-Seg-M	1.59±0.48	7.52±3.78	.927±.017	1.91×10^6
	ACNN-Seg	1.37±0.42	7.89±3.83	.939±.017	1.60×10^6
	p-values	$p \ll 0.001$	$p \approx 0.890$	$p \ll 0.001$	-
<i>Myo-cardium</i>	2D-FCN [264]	1.58±0.44	9.19±7.22	.727±.046	1.39×10^6
	3D-Seg	1.48±0.51	10.15±10.58	.773±.038	1.60×10^6
	3D-UNet [49]	1.45±0.47	9.81±11.77	.764±.045	1.64×10^6
	AE-Seg [217]	1.51±0.29	8.52±2.72	.779±.033	1.68×10^6
	3D-Seg-MAug	1.37±0.41	9.41±9.17	.785±.041	1.60×10^6
	AE-Seg-M	1.32±0.26	7.12±2.79	.791±.036	1.91×10^6
	ACNN-Seg	1.14±0.22	7.31±3.59	.811±.027	1.60×10^6
	p-values	$p \ll 0.001$	$p \approx 0.071$	$p \ll 0.001$	-

however, additional convolutional kernels are used in the AE as suggested in [217]. To observe the influence of the AE model’s capacity on the AE-Seg model’s performance, we performed experiments using different number of AE pars, and the largest capacity case is denoted by AE-Seg-M.

The results of the experiments are provided in Table 6.1 together with the capacity of each model. Statistical significance of the results is verified by performing the Wilcoxon signed-rank test between the top two performing methods for each evaluation metric. Based on these results we can draw three main conclusions: (I) Slice by slice analysis [6, 264] significantly under-performs compared to the proposed sub-pixel and ACNN-Seg segmentation methods. In particular, the dice score metrics are observed to be lower since 2D analysis can yield poor

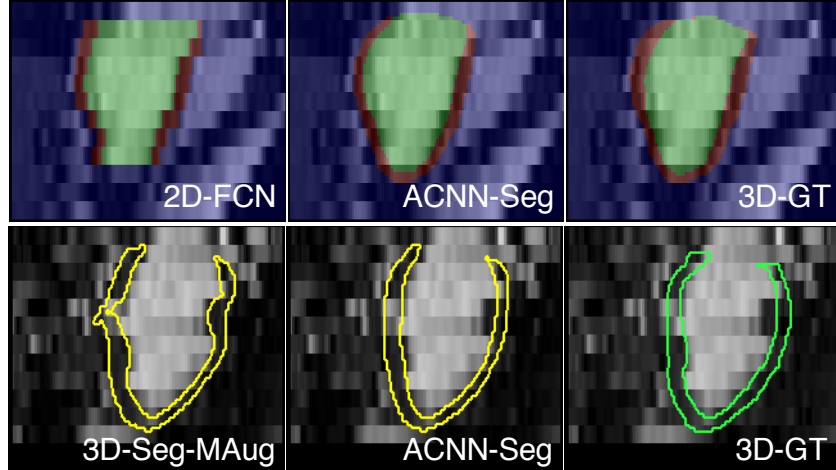


Figure 6.7: Segmentation results on two different 2D stack cardiac MR images. The proposed ACNN model is insensitive to slice misalignments as it is anatomically constrained and it makes less errors in basal and apical slices compared to the 2D-FCN approach. The results generated from low resolution image is better correlated with the HR ground-truth annotations (green).

performance in basal and apical parts of the heart as shown in Figure 6.7. Previous slice by slice segmentation approaches validated their methods on LR annotations; however, we see that the produced label maps are far off from the true underlying ventricular geometry and it can be a limiting factor for the analysis of ventricle morphology. Similar results were obtained in clinical studies [174], which however required HR image acquisition techniques. (II) The results also show that introduction of shape priors in segmentation models can be useful to tackle false-positive detections and motion-artefacts. As can be seen in the bottom row of Figure 6.7, without the learnt shape priors, label map predictions are more prone to imaging artefacts. Indeed, it is the main reason why we observe such a large difference in terms of Hausdorff distance. For endocardium labels, on the other hand, the difference in dice score metric is observed to be less due to the larger size of the LV blood pool compared to the myocardium.

Lastly (III), we observe a performance difference between the cascaded AE based segmentation (AE-Seg [217]) and the proposed ACNN-Seg models: the segmentations generated with the former model are strongly regularised due to the second stage AE. It results in reduced Hausdorff distance with marginal statistical significance, but the model overlooks fine details of the myocardium surface since the segmentations are generated only from the coarse level

feature-maps. More importantly, cascaded approaches add additional computational complexity due to the increased number of filters, which could be redundant given that the standard segmentation model is able to capture shape properties of the organs as long as it has a large receptive field and is optimised with shape constraints. In other words, shape constraints can be learnt and utilised in standard segmentation models, as shown in ACNN-Seg, without a need for additional model parameters and computational complexity. We also analysed the performance change in AE-Seg with respect to the number of parameters, which shows that the small capacity AE-Seg model (8×10^4 pars) is not suitable for cardiac image segmentation as the second stage in the cascaded model does not improve the performance significantly.

We performed additional segmentation experiments using only the T-L network. In detail, the input LR image is passed first through the predictor network and then the extracted codes are fed to the decoder network shown in Figure 6.3. Label map predictions are collected at the output of the decoder and they are compared with the same ground-truth annotations described previously, which was similar to the AE based segmentation method proposed in [6, 7]. We observed that reconstruction of label-maps from low dimensional representations was limited since the ventricle boundaries were not delineated properly but rather a rough segmentation was generated (DSC: .734). We believe that this is probably the main reason why the authors of [6] proposed the use of a separate deformable model at the output of a NN. Nevertheless, the proposed ACNN-Seg does not require an additional post-processing step.

6.3.4 Cardiac 3D Ultrasound Image Segmentation

In the second experiment, the proposed model is evaluated on 3D cardiac ultrasound data which is described in Sec 6.3.1. Segmentation models are used to delineate endocardial boundaries and the segmentations obtained on ED and ES frames are later used to measure volumetric indices such as ejection fraction (EF). The models are compared also in terms of surface to surface distance errors of their corresponding endocardium segmentations. As a baseline CNN method, we utilised the fully convolutional network model suggested by [45] for multi-view

Table 6.2: 3D-US cardiac image sequences (in total 30) are segmented into LV cavity and background. Segmentation accuracy is evaluated in terms of Dice score (DSC), surface-to-surface distances. The consistency of delineations on both ED and ES phases are measured in terms computed ejection fraction (EF) values. The proposed ACNN-Seg method is compared against state-of-the art deformable shape fitting [14] and fully-convolutional 3D segmentation [45] methods.

		Mean Dist. (mm)	Hausdorff Dist. (mm)	Dice Score (%)	Ejection Fraction (Corr) (Bias+LOA) (ml)	
<i>End - diastole</i>	BEAS [14]	2.26±0.73	8.10±2.66	.894±.041	0.889	-6.78±27.71
	FCN [45]	1.98±1.03	11.94±9.46	.906±.026	0.885	2.74±12.01
	ACNN-Seg	1.89±0.51	6.96±1.75	.912±.023	0.913	1.78±10.09
<i>End - systole</i>	BEAS [14]	2.43±0.91	8.13±3.08	.856±.057	-	-
	FCN [45]	2.83±1.89	12.45±10.69	.872±.050	-	-
	ACNN-Seg	2.09±0.77	7.75±2.65	.873±.051	-	-
p-values		$p < 0.01$	$p < 0.001$	$p \approx 0.05$	-	-

3D-US image segmentation problem. It is also observed to be more memory efficient compared to the standard 3D-UNet architecture [49]. Additionally, we compare our proposed model against the CETUS'14 challenge winner approach (BEAS) [14] that utilised deformable models to segment the left ventricular cavity. The challenge results can be found in [20].

The experimental results, given in Table 6.2, show that neural network models outperforms previous state-of-the-art approaches in this public benchmark dataset although the training data size was limited to 15 image sequences. The experimental results were evaluated in a blinded fashion by uploading the generated segmentations from separate 30 sequences into the CETUS web platform. The main contribution of ACNN model over the standard FCN approaches is the improved shape delineation of the LV, as it can be seen in terms of the distance error metrics. In particular, Hausdorff distances were reduced significantly as global regularisation reduces the amount of spurious false positive predictions and enforces abnormal LV segmentation shapes to fit into the learnt representation model. This situation is illustrated in Figure 6.8. Similarly, we observed an improvement in terms of normalised Dice score, which was quantitatively not significant due to large volumetric size of the LV cavity. Lastly, we compared the extracted ejection fraction results to understand both the accuracy of segmentations and also the consistency of these predictions on both ED and ES phases. It is observed that the ACNN approach

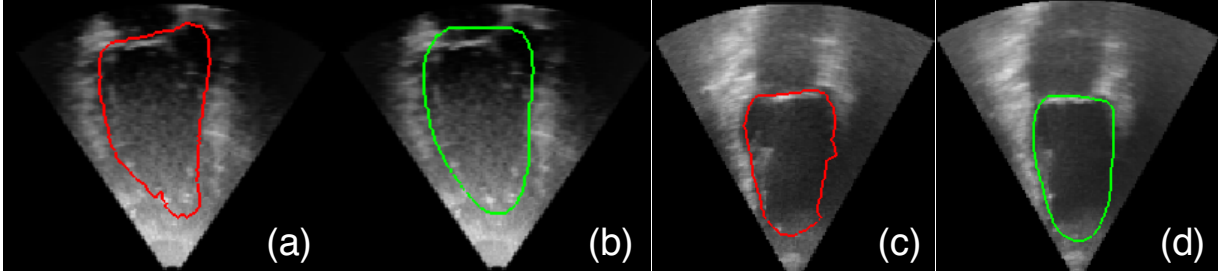


Figure 6.8: (a) Cavity noise limits accurate delineation of the LV cavity in apical areas. (b) The segmentation model can be guided through learnt shape priors to output anatomically correct delineations. (c) Similarly, it can make accurate predictions even when the ventricle boundaries are occluded.

ensures better consistency between frames although none of the methods have used temporal information.

The reported results could be further improved by segmenting both ED and ES frames simultaneously or by extracting the temporal content from the sequences. For instance, propagation of ED masks to ES frames through optical flow has been shown to be a promising way to achieve this goal. However, this study mainly focuses on demonstrating the advantages of using priors in neural network models, and achieving the best possible segmentation accuracy was not our main focus.

6.3.5 Cardiac MR Image Enhancement

The proposed ACNN model is also applied to the image SR problem and compared against the state-of-the-art CNN model used in medical imaging [191]. The cardiac MR dataset, described in Sec. 6.3.1, was split into two disjoint subsets: training (1000) and testing (200). At testing time, we evaluated our model with both LR-HR clinical image data. In training, however, LR images are synthetically generated from clinical HR data using the MR acquisition model discussed in [103]. More details about the acquisition model can be found in [191].

The quality of the upsampled images is evaluated in terms of SSIM metric [269] between the clinical HR image data and reconstructed HR images. SSIM measure assesses the correlation

Table 6.3: Average inference time (Inf-T) of the SR models per input LR image (120x120x12) using a GPU (GTX-1080). ACNN-SR and SR-CNN [191] models are given the same number of filters and capacity. MOS [141] results, received from the clinicians (R1 and R2), are reported separately.

	SSIM [269]	MOS-R1	MOS-R2	Inf-T
Linear	.777±.043	2.71±0.82	2.60±.91	-
B-Spline	.779±.053	2.77±0.89	2.64±.84	-
SR-CNN [191]	.783±.046	3.59±1.05	3.85±.70	.29 s
3D-UNet [49]	.784±.045	3.55±0.92	3.99±.71	.07 s
ACNN-SR	.796±.041	4.36±0.62	4.25±.68	.06 s
p-values	$p \ll 0.001$	$p < 0.001$	$p < 0.01$	-

of local structures and is less sensitive to image noise than PSNR which is not used in our experiments since small misalignments between LR-HR image pairs could introduce large errors in the evaluation due to pixel by pixel comparisons. More importantly, intensity statistics of the images are observed to be different for this reason PSNR measurements would not be accurate. In addition to the SSIM metric, we used the mean opinion score (MOS) testing [141] to quantify the quality and similarity of the synthesised and real HR cardiac images. Two expert cardiologists were asked to rate the upsampled images from 1 (very poor) to 5 (excellent) based on the accuracy of the reconstructed LV boundary and geometry. To serve as a reference, the corresponding clinical LR and HR images are displayed together with the upsampled images that are anonymised for a fair comparison.

In Table 6.3, SSIM and MOS scores for the standard interpolation techniques, SR-CNN, and the proposed ACNN-SR models are provided. In addition to the increased image quality, the ACNN-SR model is computationally more efficient in terms of run-time in comparison to the SR-CNN model [191] by a factor of 5. This is due to the fact that ACNN-SR performs feature extraction in the low dimensional image space. Furthermore, we investigated the contribution of shape regularisation term in the application of SR, which is visualised in Figure 6.9.

Moreover, we investigated the use of SR as a pre-processing technique for subsequent analysis such as image segmentation, similar to the experiments reported in [191]. In that regard, the proposed SR model and U-Net segmentation models are concatenated to obtain HR seg-

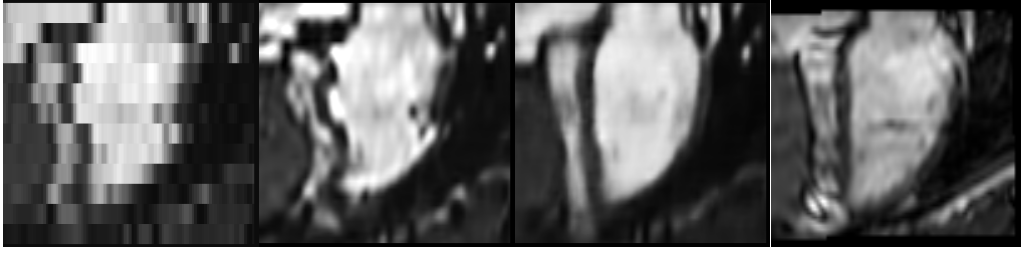


Figure 6.9: Image super-resolution (SR) results. From left to right, input low resolution MR image, baseline SR approach [191] (no global loss), the proposed anatomically constrained SR model, and the ground-truth high resolution acquisition.

mentation results. However, we observed that the proposed baseline sub-pixel segmentation model (3D-Seg), which merges both SR and segmentation tasks, performs better than the concatenated models. The 3D-Seg approach uses the convolution kernels more efficiently without requiring the model to output a high-dimensional intensity image. For this reason, SR models should be trained by taking into account the final goal and in some cases it's not required to reconstruct a HR intensity image for HR analysis.

6.3.6 Learnt Latent Representations and Pathology Classification

The jointly trained T-L model and its latent representations are analysed and evaluated in the experiment of image pathology classification. This experiment focuses on understanding the information stored in the latent space and also investigates whether they can be used to distinguish healthy subjects from dilated and hypertrophic cardiomyopathy patients. For this, we collected 64 dimensional codes from segmentation images of the cardiac MR dataset explained in Sec. 6.3.1. Similarly, principal component analysis (PCA) was applied to the same segmentation images (containing LV blood-pool and myocardium labels) to generate 64 dimensional linear projection of the labels, which requires additional spatial-normalisation prior to linear mapping. The generated codes were then used as features to train an ensemble of decision trees to categorise each image. We used 10-fold cross-validation on 60 CMR sequences and obtained 76.6% vs 83.3% accuracy using PCA and T-L codes extracted from ED phase. By including the codes from ES phase, the classification accuracies were improved to 86.6%

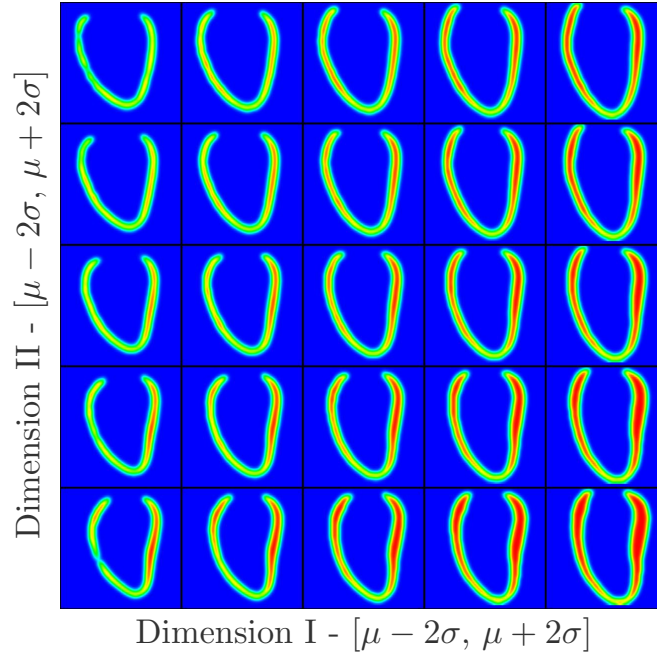


Figure 6.10: Anatomical variations captured by the latent representations in T-L network (swipe from $\mu - 2\sigma$ to $\mu + 2\sigma$). Based on our observation, the first and second dimensions capture the variation in the wall thickness of the myocardium (x-axis) and lateral wall of the ventricle (y-axis).

vs 91.6%. This result shows that although the AE and T-L models are not trained with the classification objective, they can still capture anatomical shape variations that are linked to cardiac related pathologies. In particular, we observed that some latent dimensions are more commonly used than others in tree node splits. By sampling codes from the latent space across these dimensions, we observed that the network captures the variation in wall thickness and blood pool size as shown in Figure 6.10. Since we obtain a regular and smooth latent representation, it is possible to transverse along the latent space and generate LV shapes by interpolating between data points. It is important to note that classification accuracies can be further improved by training the AE and T-L models with a classification objective. Our main goal in this experiment was to understand whether the enforced prior distributions contain anatomical information or they are abstract representations only meaningful to the decoder of the AE.

6.4 Discussion and Conclusion

In this work, we presented a new image analysis framework that utilises autoencoder (AE) and T-L networks as regularisers to train neural network (NN) models. With this new training objective, at testing time NNs make predictions that are in agreement with the learnt shape models of the underlying anatomy, which are referred as image priors. The experimental results show that the state-of-the-art NN models can benefit from the learnt priors in cases where the images are corrupted and contain artefacts. The proposed regulariser model can be seen as an application-specific training objective. In that regard, our model differentiates from the VGG-Net [246] feature based training objectives [120, 141]. VGG features tend to be more general purpose representations that are learnt from ImageNet dataset containing natural images of a large variety of objects. In contrast to this, our AE model is trained solely on cardiac segmentation masks and features are customised to identify anatomical variations observed in the heart chambers. For this reason, we would expect the AE features of the segmentations to be more distinctive and informative.

As an alternative to the proposed framework, label space dependencies could be exploited also through adversarial loss (AL) objective functions. Such approaches have been used successfully in natural image super-resolution (SR) [141] and segmentation [164] tasks. In SR application, AL enables the SR network to hallucinate fine texture detail, and the synthesized HR images appear qualitatively more realistic. However, at the same time the PSNR and SSIM scores are usually worse. For this reason, the authors of [141] have pointed out that adversarial training may not be suitable for medical applications, where the accuracy and fidelity of the visual content more important than the qualitative appearance of the HR images. Moreover, we believe that adversarial training comes at the expense of less interpretability of the regularisation term and unstable model training behaviour, which still remains an open research problem.

Additionally, in the experiments we demonstrated that the learnt codes can be used as biomarkers for classification of cardiac related pathologies and we analysed the distribution of the learnt

latent space. This latent space can be further constrained to be Gaussian distributed by replacing the proposed regularisation model with a variational autoencoder. However, this design choice was not considered in our ACNN framework due to two main reasons: (I) the additional K-L divergence term (constraint) would reduce the representation power of the AE; thus, the local anatomical variations would not be captured in detail. (II) A generative AE model is not essential for the regularisation of the proposed segmentation and SR models. A variational architecture would be useful if it was required to sample random instances from the latent space and reconstruct anatomically meaningful segmentation masks; however, in our framework we are only interested in the anatomy specific AE features for model regularisation.

The presented ACNN framework is not only limited to the medical image segmentation and SR tasks but can be extended to other image analysis tasks where prior knowledge can provide model guidance and robustness. In that regard, future research will focus on the application of ACNN to the problems such as human pose estimation, anatomical and facial landmark localisation on partially occluded image data.

6.5 Implementation Details

In this section, we give the details about the meta-parameter choices and sampling strategies. In both the super-resolution (SR) and segmentation (Seg) models, the weight decay regularisation term is weighted by $\lambda_2 = 5 \times 10^{-6}$, and gradient descent learning-rate is fixed to $lr = 0.001$. The weight of global priors is chosen experimentally to be $\lambda_1 = 0.01$. Mini batch-size, which is the number of samples used for each back-propagated gradient update, is set to be 8 samples. The models are trained with full images without a need for patch extraction since the cardiac 2D MR image stack size is relatively smaller compared to the available GPU memory. In the SR problem, the through-plane upsampling factor is fixed to $K = 5$ and the synthetic low-resolution training samples are generated simply by filtering high-resolution images with a Gaussian blurring kernel ($\sigma = 4.0$) along the through plane direction. The blurring operation

Table 6.4: Structure of the predictor model: The model maps the input HR intensity image (120x120x60) to the latent space and generates a 64-dim representation. The size, number, and stride of the learnt convolution (Conv) kernels are provided. The filters operate on different image scales ($S1$ - $S4$) and each convolution operation is followed by a non-linear unit (ReLU).

		Size	Stride	# Kernels	Non-linearity
$S1$	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S2$	Conv	(f:3,3,3)	(s:2,2,2)	(N:64)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S3$	Conv	(f:3,3,3)	(s:2,2,2)	(N:128)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:128)	ReLU
$S4$	Conv	(f:3,3,3)	(s:2,2,2)	(N:256)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:1)	ReLU
	FC	-	-	(N:64)	None

is followed by a decimation operator along the same image dimension. In the segmentation problem, the Sorensen-Dice loss was tested in the experiments as an alternative to the cross-entropy loss, yet we observed a degraded performance since the latter is a smoother function.

6.5.1 Network Structure

In this section, we give the network structures of the autoencoder (AE) and the predictor, which together build the T-L network. The details are provided in Table 6.4 and 6.5. As can be seen in the tables, residual connections are not used in our models (10-18 layers) since they do not provide significant accuracy gains for smaller networks as reported in [141]. Additionally, non-linear layers are not applied on the lower dimensional latent representations since that would further constrain the autoencoder. The number of the hidden units (64) is chosen experimentally, and optimal configurations can be explored to improve the reconstruction performance. The input layer of the AE model is extended to multi-label segmentation maps through one-hot image representation, where each label is converted into a separate channel at the input. Lastly, in both models (AE and PR) each convolution layer, except the last one, is followed by batch-normalisation for better convergence behaviour.

Table 6.5: Structure of the autoencoder (AE) model: The encoder part maps the given input segmentation map (120x120x60) to the latent space through convolution (Conv) and fully-connected (FC) layers. The decoder part recovers the input from the low-dimensional representation and outputs a segmentation map (120x120x60). The size, number, and stride of the learnt convolution (Conv) kernels are provided. The filters operate on different image scales ($S1$ - $S4$) and each convolution operation is followed by a non-linear unit (ReLU).

		Kernel	Stride	# Kernels	NonLin
$S1$	Conv	(f:3,3,3)	(s:2,2,1)	(N:16)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:16)	ReLU
$S2$	Conv	(f:3,3,3)	(s:2,2,2)	(N:32)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S3$	Conv	(f:3,3,3)	(s:2,2,2)	(N:64)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S4$	Conv	(f:3,3,3)	(s:3,3,3)	(N:1)	ReLU
HC	FC	-	-	(N:64)	None
	FC	-	-	(N:125)	ReLU
$S4$	Deconv	(f:7,7,7)	(s:3,3,3)	(N:64)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S3$	Deconv	(f:4,4,4)	(s:2,2,2)	(N:32)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S2$	Deconv	(f:4,4,4)	(s:2,2,2)	(N:16)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:16)	ReLU
$S1$	Deconv	(f:4,4,1)	(s:2,2,1)	(N:16)	ReLU
	Conv	(f:3,3,3)	(s:1,1,1)	(N:3)	None

Table 6.6: Structure of the segmentation and super-resolution models. The model takes a low-resolution image as an input (120x120x60) and performs a sub-pixel classification or regression. The size, number, and stride of the learnt convolution (Conv) and deconvolution (Deconv) kernels are provided. The filters operate on different image scales ($S1$ - $S5$) and each convolution operation is followed by a non-linear unit (ReLU).

	Layer	Kernel	Stride	# Kernels	NonLin
$S1$	Conv_1_1	(f:3,3,3)	(s:1,1,1)	(N:16)	ReLU
	Conv_1_2	(f:3,3,3)	(s:1,1,1)	(N:16)	ReLU
$S2$	Conv_2_1	(f:3,3,3)	(s:2,2,1)	(N:32)	ReLU
	Conv_2_2	(f:3,3,3)	(s:1,1,1)	(N:32)	ReLU
$S3$	Conv_3_1	(f:3,3,3)	(s:2,2,1)	(N:64)	ReLU
	Conv_3_2	(f:3,3,3)	(s:1,1,1)	(N:64)	ReLU
$S4$	Conv_4_1	(f:3,3,1)	(s:2,2,1)	(N:64)	ReLU
	Conv_4_2	(f:3,3,1)	(s:1,1,1)	(N:64)	ReLU
$S5$	Conv_5_1	(f:3,3,1)	(s:3,3,1)	(N:128)	ReLU
	Deconv_5_1	(f:6,6,1)	(s:3,3,1)	(N:128)	ReLU
$S4$	Concatenate	Deconv_5_1 and Conv_4_2			None
	Conv_4_3	(f:3,3,1)	(s:1,1,1)	(N:64)	ReLU
	Deconv_4_1	(f:4,4,1)	(s:2,2,1)	(N:64)	ReLU
$S3$	Concatenate	Deconv_4_1 and Conv_3_2			None
	Conv_3_3	(f:3,3,1)	(s:1,1,1)	(N:64)	ReLU
	Deconv_3_1	(f:4,4,1)	(s:2,2,1)	(N:64)	ReLU
$S2$	Concatenate	Deconv_3_1 and Conv_2_2			None
	Conv_2_3	(f:3,3,1)	(s:1,1,1)	(N:32)	ReLU
	Deconv_2_1	(f:4,4,1)	(s:2,2,1)	(N:32)	ReLU
$S1$	Concatenate	Deconv_2_1 and Conv_1_2			None
	Conv_1_3	(f:3,3,1)	(s:1,1,1)	(N:16)	ReLU
	Deconv_1_1	(f:1,1,11)	(s:1,1,5)	(N:64)	ReLU
	Conv_1_4	(f:3,3,3)	(s:1,1,1)	(N:3)	None

Chapter 7

Summary and Future Work

In this chapter, we first give a brief overview of the thesis' scope and the research problems. In addition, the proposed and achieved technical contributions are shortly summarised item by item. Furthermore, we discuss about the current limitations of the proposed algorithms and propose possible directions for future research. Lastly, we conclude the chapter with a general outlook on the thesis and possible future cardiac imaging research.

7.1 Summary

Diagnosis and treatment of patients with suspected or known cardiovascular disease has become increasingly dependent on the use of non-invasive cardiac imaging techniques. Two of the well established imaging modalities, ultrasound (US) and magnetic resonance imaging (MRI), allow accurate assessment of heart function and morphology by precise delineation of the myocardial anatomy with high tissue contrast and resolution [51, 124]. Moreover, image based quantitative analysis has been increasingly recognised as crucial [231] for more detailed and objective analysis of the heart anatomy, which involves derivation of physiological clinical indices from the imaging data. Particular examples are ventricular mass and volumetric measurements, myocardial wall thickness and strain analysis.

Image analysis tools have an important role in accurate and objective assessment of these clinical measurements from multi-modal imaging data. However, full automation and reliable utilisation of these software tools still face technical challenges due to limitations imposed by the imaging modalities:

- Imaging data artefacts that are specific to the employed imaging modality. Common examples are patient respiratory motion (CMR), signal drop-out and shadowing (US).
- Low resolution image data in the through-plane direction, which is commonly observed in 2D cine cardiac MR images.
- Differences in appearance of the left ventricular anatomy, which is observed between different imaging modality types such as US and MRI.

The thesis has focused on studying the adverse effects of these limitations imposed by the imaging modalities on common automated analysis tasks such as image registration, segmentation, and image super-resolution. The experimental results, provided in the previous chapters, show that the proposed frameworks can lead to more reliable solutions by overcoming the limitations observed in cardiac imaging.

From a general perspective, the presented approaches share two common features: (I) machine learning models and a large number of training datasets are used for better understanding of the image data, and (II) a new image representation is extracted from raw intensities to provide additional information for subsequent image analysis. The former feature brings additional domain knowledge into the image processing framework, which is later used to guide the analysis and reduces the algorithm sensitivity. For instance, high resolution appearance of the left ventricle, which is learnt from the training samples, can inform and guide the segmentation modules to better delineate the anatomical boundaries as shown in Chapters 5 and 6. Learning new image representations, on the other hand, provide better data interpretability and task simplification in image analysis. To elaborate on that we can consider the multi-modality

comparison experiments given in Section 3.6, where US and MR cardiac images were compared with each other in the learnt PEM space. By extracting the endocardial surface boundaries, we have been able to compare the properties of each modality in terms of how they visualise the left ventricle. Another example to consider is the multi-modal cardiac image registration problem discussed in Section 3.3. The learnt PEM space enables us to perform the spatial alignment between images only with respect to the myocardial boundaries, where the background intensity data dissimilar between US and MRI is not involved in the optimisation function. In this way, the registration task is simplified. The next section summarises the other achievements presented in this thesis, and gives examples of how the learnt representations could provide us a better understanding of cardiac images.

7.1.1 Achievements

Multi-modal cardiac image registration: In Chapter 3, we have investigated feature based intra- and inter-modality cardiac image registration problems: multi-atlas 3D-US segmentation, and registration of pre- and intra-operative MR/US cardiac images. An analysis on existing approaches which rely on feature maps, has shown that the existing methods such as MIND [109] and local phase [131] descriptors show large sensitivity to US image quality, and the registration capture range is usually quite limited. In other words, the initialisation of the registration algorithm is quite crucial for a reliable operation. To address these problems, we have first introduced a new image feature type namely probabilistic edge maps (PEMs). In the context of this thesis, it is referred as an image representation type. Later, we have demonstrated its application to multi-modal US/MR registration and 3D-US multi-atlas segmentation problems, where the latter involves intra-subject US registration as well. Besides the improved registration accuracy and increased capture range, the work has made the following two contributions as well: (I) This was one of the earliest works that has utilised a supervised machine learning model within a multi-modal medical image registration framework, (II) since PEMs outline only the organ surface boundary, a target organ-specific image similarity metric

can be defined which significantly simplifies the multi-modal registration problem.

In the multi-atlas (MA) segmentation problem, we have demonstrated that the PEM-MA framework achieves top segmentation accuracy in the CETUS challenge, which was held as part of the MICCAI 2014 conference. Based on these two examples, we can conclude that the PEM representation is a good alternative to intensity based analysis techniques when 3D-US cardiac images are involved in the processing pipeline. However, as a downside, the proposed approach and PEM representation do require manual surface annotations for model training.

At end of the chapter, we have presented a comparison study between cardiac US and cine MR images. The objective was to understand the underlying reasons of why the LV volumetric measurements performed on 3D-US image data underestimate the values obtained from the cardiac MR images. To best of our knowledge, earlier clinical studies [55, 170] have not investigated this problem on a local image pixel level. In contrast, we have mapped the spatially aligned endocardium segmentations into a common space in form of a cardiac atlas, and analysed the local differences between the US/MR segmentations locally. In that regard, future studies can benefit from PEM based registration method and atlas based analysis techniques to investigate this problem in more depth.

Anatomical landmark localisation in cardiac images: In Chapter 4, a new methodology is proposed to localise anatomical landmarks in cardiac MR images, which can find several application areas such as image acquisition planning, or image registration initialisation. In particular, the second feature is an important requirement for most atlas based segmentation techniques [11]. As an extension to the PEM representation, we have demonstrated that extracted surface points can also help to identify and localise key anatomical landmarks (LM) such as RV insert points and apex. This goal is achieved by allowing the surface points to cast votes for each individual LM location, and the collected votes are later aggregated to obtain the final result. In this respect, the propose framework is similar to the Hough voting [90], but in our case the information is collected only from the myocardium surface points via PEMs.

Moreover, we have proposed a new node split model that could effectively learn population specific distributions for classification and regression tasks. The experimental results show the improved accuracy over the standard classification and Hough voting based landmark localisation techniques.

Cardiac cine-MR image quality enhancement: Cine cardiac MR (CMR) images are often acquired in stacks of 2D slices over several heartbeat cycles and patients are required to hold their breath repeatedly for the acquisition of each slice. Since this is the common protocol in clinical practice, the quantitative analysis suffer from low resolution in the direction orthogonal to the slice acquisition plane. This technical problem has been studied over the years to achieve faster and also accurate image acquisition techniques such as parallel imaging and partial k-space sampling [166, 211]. In parallel to these attempts, super-resolution techniques have been developed and applied to cine cardiac MR images such as atlas [243] and regression model [2] based approaches. However, their application have been usually impractical either due to long computation time or insufficient model accuracy.

To address these challenges, we have proposed a CNN based super-resolution approach (SR-CNN) in Chapter 5. SR-CNN significantly improves the performance in terms of both computation time and accuracy by learning more complex regression models. In the experiments, we have demonstrated that subsequent image analysis such as motion tracking or LV segmentation can benefit from super-resolution when it is used a pre-processing technique to enhance the resolution of the input CMR images. Lastly, it is worth to mention that in Section 3.5, we have proposed another image enhancement technique for CMR images. The presented framework utilises PEMs to reduce the respiratory motion artefacts observed between adjacent slices in image stacks.

Integration of prior anatomical knowledge in CNNs Motion artefacts observed in image data usually degrade the performance of the segmentation and super-resolution CNN models that are proposed in the previous chapter. Similarly, the same observation can be made in

3D-US data when the ventricle boundaries are missing due to shadowing. Instead of using an external ad-hoc pre-processing technique to correct these artefacts, we have explored combined pre-processing and analysis techniques within the CNN framework. In other words, the trained CNN models are taught to be aware of the underlying LV geometry, shape and anatomy. By achieving this goal, one could obtain a multi-atlas equivalent of CNN segmentation and super-resolution models.

We incorporated this prior knowledge by training the standard baseline models with a shape regularisation network which is learnt separately using segmentation labels only. The experiments show that the new architecture, namely anatomically constrained neural networks (ACNN), can significantly boost the performance of baseline approaches when the testing data is not self-sufficient for the analysis due to imaging artefacts.

7.2 Limitations and Future Work

In this section, the main limitations of the presented image analysis methods are listed together with some suggestions to tackle these limitations in future research. It is important to note that the technical drawbacks are not limited to the items listed below:

Better ways to encode the structural information: The proposed PEM representation is mainly intended to capture the structural information in 3D-US images. Nevertheless, one dimensional representation for each pixel point does not necessarily contain all the required information to guide an image registration algorithm. For instance, the representation could be enriched by including boundary angle and polarity information. Such multi-channel representations can be used in image registration objective functions as in the case of [148]. Moreover, the representation power of the features can be further enhanced by exploring new feature spaces through CNN models. One particular example could be adversarial training of MR/US pairs to obtain a modality independent feature space.

Learning a PEM model without training data: One of the main limitations of the PEM based registration technique is the requirement of training data. Additional information given by the external training data provided algorithm robustness, but its applicability is usually limited by the availability of the data. To circumvent this problem, we believe that future research could investigate domains such domain knowledge transfer or weakly supervised learning. A particular example was proposed in [149] to learn spatio-temporal object boundaries from motion edges in an unsupervised way.

Optimal ways to register PEMs: In the proposed method, the edge maps obtained from US and MR images are registered using a robust block matching algorithm [197] that is designed for intensity images. Although with the current setting this approach provides good spatial alignment results, the performance could be further enhanced by exploring the use of robust surface alignment techniques or point matching algorithms [98, 181].

Adversarial learning should be used in the proposed CNN models: Recently, adversarial learning (AL) has been shown to be a powerful technique to design a customised loss function that could not be formulated explicitly in a closed-form expression. Recent studies [101, 141] have shown that models trained with AL objective functions can capture the true underlying data distribution, which could enable us to capture the data manifold of semantic segmentations or the high-resolution cardiac image representations. In other words, the performance of ACNN models can be further improved by learning the shape space through the AL objective function. However, stable model training behaviour is still an open research problem in AL, and there is no well-established strategy to train models with this objective. Future research should investigate this problem and ACNN training may no longer be required to obtain 3D LV mesh representations from 2D stack CMR images.

Optimisation of multiple loss terms in CNN training is inefficient: The proposed ACNN model is trained by minimising an objective function that is composed of shape prior

and reconstruction loss terms. The two loss functions are coupled with each other through a weighting term, and the optimal value for this parameter is computed through a grid-search approach. A similar problem can be encountered in optimisation of both reconstruction and AL loss terms. In [122], a dynamic loss weighting scheme has been proposed to obtain a better convergence behaviour. However, there has not been a proposed systematic way to solve this problem and we believe that future research should investigate better ways to couple different loss terms. One possible direction is combining multiple objectives through uncertainty weighting as suggested in [125], which allows to combine quantities with different scales in both classification and regression settings.

Super-resolution should not be the end-goal: The thesis has presented two different CNN models for the cardiac image super-resolution (SR) problem. Many other studies have explored this research area and proposed different SR solutions (e.g. coupled dictionary learning or atlas based techniques). Medical image SR in general can potentially provide better qualitative assessment and guidance for subsequent analysis, as it is shown in Chapter 5. However, with the advent of CNNs, it is no longer required to reconstruct a high resolution (HR) intensity image to perform HR volumetric measurements. In more detail, end-to-end training of 2D stack data and high-resolution segmentations enables this feature. Moreover, models trained with this objective function significantly outperform the concatenated SR and segmentation models as shown in Chapter 6. This is mainly due to the fact that it is no longer required to spare the model capacity to reconstruct HR intensity images. For this reason, SR may not be required for HR image analysis and the SR training objective function needs to be designed by considering the end-goal of the analysis framework.

List of Publications

Book chapters

- C. Baumgartner, **O. Oktay**, D. Rueckert. “Fully Convolutional Networks in Medical Imaging: Applications to Image Enhancement and Recognition”. Deep Learning and Convolutional Neural Networks for Medical Image Computing. Edited by Lu, L., Zheng, Y., Carneiro, G., Yang, L. . Springer, 2017.

Journal publications

- **O. Oktay**, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. Cook, A. Marvao, T. Dawes, D. O'Regan, B. Kainz, B. Glocker, D. Rueckert. “Anatomically Constrained Neural Networks (ACNN): Application to Cardiac Image Enhancement and Segmentation”. IEEE Transactions on Medical Imaging (IEEE TMI), In print, Aug 2017.
- A. Gomez, **O. Oktay**, D. Rueckert, G. P. Penney, J. A. Schnabel, J. M. Simpson and K. Pushparajah. “Regional Differences in End-Diastolic Volumes between 3D Echo and CMR in HLHS Patients”. Frontiers in Pediatrics, vol. 4, Dec 2016.
- **O. Oktay**, W. Bai, M. Rajchl, M. P. Heinrich, B. Glocker, D. Rueckert. “Stratified Decision Forests for Accurate Anatomical Landmark Localization”. IEEE Transactions on Medical Imaging (IEEE TMI) 36.1 (2017) pp: 332-42.

- M. Rajchl, M. Lee, **O. Oktay**, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. Rutherford, J. Hajnal, B. Kainz, D. Rueckert. “DeepCut: Object Segmentation from Bounding Box Annotations using Convolutional Neural Networks”. *IEEE Transactions on Medical Imaging (IEEE TMI)* 36.2 (2017) pp: 674-683.

Conference publications

- W. Bai, **O. Oktay**, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, D. Rueckert. “Semi-Supervised Learning for Network-Based Cardiac MR Image Segmentation”. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. September 2017. [AR 27%]
- M. Heinrich, **O. Oktay**. “BRIEFnet: Deep Pancreas Segmentation using Sparse Dilated Convolutions”. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. September 2017. [AR 27%]
- G. Tarroni, **O. Oktay**, W. Bai, A. Schuh, H. Suzuki, J. Palmbach, B. Glocker, D. Rueckert. “Learning-Based Heart Coverage Estimation for Short-Axis Cine Cardiac MR Images”. *Functional Imaging and Modeling of the Heart (FIMH)*. Springer, 2017. 73-82
- G. Tarroni, **O. Oktay**, A. Schuh, W. Bai, A. de Marvao, D. O’Regan, S. Cook, D. Rueckert. “Slice Realignment for Motion-Corrupted Stacks of Short-Axis Cine Cardiac MR Images based on 3D Probabilistic Edge Maps”. *Proceedings of the 25th International Society for Magnetic Resonance in Medicine (ISMRM) Annual Meeting and Exhibition, Hawaii, USA, April 2017*.
- M. P Heinrich, **O. Oktay**. “Accurate Intervertebral Disc Localisation and Segmentation in MRI using Vantage Point Hough Forests and Multi-Atlas Fusion”. *Proceedings of*

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.

the MICCAI Workshop on Computational Methods and Clinical Applications for Spine Imaging (CSI), Springer, 2016.

- **O. Oktay**, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, Antonio de Marvao, S. Cook, D. O'Regan, D. Rueckert. "Multi-Input Cardiac Image Super-Resolution using Convolutional Neural Networks". Medical Image Computing and Computer-Assisted Intervention (MICCAI) Conference 2016. [AR 28%]
- **O. Oktay**, G. Tarroni, W. Bai, A. de Marvao, D. O'Regan, S. Cook, D. Rueckert. "Respiratory motion correction for 2D cine cardiac MR images using probabilistic edge maps". In Computing in Cardiology Conference (CinC), September 2016.
- W. Bai, **O. Oktay**, D. Rueckert. "Classification of Myocardial Infarcted Patients by Combining Shape and Motion Features". Statistical Atlases and Computational Models of the Heart (STACOM). Springer, 2015. 140-145.
- **O. Oktay**, A. Schuh, M. Rajchl, K. Keraudren, A. Gomez, M. P. Heinrich, G. Penney, D. Rueckert. "Structured Decision Forests For Multi-modal Ultrasound Image Registration". Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, 2015. 363-371. [AR 30%]
- K. Keraudren, B. Kainz, **O. Oktay**, V. Kyriakopoulou, M. Rutherford, J. V Hajnal, D. Rueckert. "Automated Localization of Fetal Organs in MRI Using Random Forests with Steerable Features". Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, 2015. 620-627. [AR 30%]
- **O. Oktay**, A. Gomez, K. Keraudren, A. Schuh, W. Bai, W. Shi, G. Penney, D. Rueckert. "Probabilistic Edge Map (PEM) for 3D Ultrasound Image Registration and Multi-atlas Left Ventricle Segmentation". Functional Imaging and Modeling of the Heart (FIMH). Springer, 2015. 223-230.
- W. Bai, D. Peressutti, **O. Oktay**, W. Shi, D. Rueckert. "Learning a Global Descriptor

of Cardiac Motion from a Large Cohort of 1000+ Normal Subjects.” *Functional Imaging and Modeling of the Heart (FIMH)*. Springer, 2015. 3-11.

- **O. Oktay**, W. Shi, J. Caballero, K. Keraudren, D. Rueckert. “Sparsity based spectral embedding: Application to multi-atlas echocardiography segmentation.” In *Proceedings of MICCAI STMI Workshop*. September 2014.
- **O. Oktay**, L. Zhang, T. Mansi, P. Mountney, P. Mewes, S. Nicolau, L. Soler, C. Chefd’hotel. “Biomechanically Driven Registration of Pre- to Intra- operative 3D Images for Laparoscopic Surgery”. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2013. 1-9. [AR 32%]

Bibliography

- [1] G. Alain and Y. Bengio, “What regularized auto-encoders learn from the data generating distribution,” *Journal of Machine Learning Research*, vol. 15, pp. 3563–3593, 2012. DOI: [abs/1211.4246](#). arXiv: [1211.4246](#).
- [2] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi, “Image quality transfer via random forest regression: Applications in diffusion MRI,” in *MICCAI*, Springer, 2014, pp. 225–232.
- [3] P. Aljabar, R. A. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert, “Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy,” *NeuroImage*, vol. 46, no. 3, pp. 726–38, 2009.
- [4] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano, “Combination strategies in multi-atlas image segmentation: Application to brain MR data,” *IEEE Trans. Med. Imag.*, pp. 1266–77, 2009.
- [5] A. J. Asman, Y. Huo, A. J. Plassard, and B. A. Landman, “Multi-atlas learner fusion: an efficient segmentation approach for large-scale data,” *Medical Image Analysis*, vol. 26, no. 1, pp. 82–91, 2015. DOI: [10.1016/j.media.2015.08.010](#).
- [6] M. R. 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*, vol. 30, pp. 108–119, 2016. DOI: [10.1016/j.media.2016.01.005](#). arXiv: [1512.07951](#).

- [7] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, *Automatic segmentation of the right ventricle from cardiac mri using a learning-based approach*, 2017. DOI: 10.1002/mrm.26631.
- [8] S. D. Babacan, R. Molina, and A. K. Katsaggelos, “Total variation super resolution using a variational approach,” in *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, IEEE, 2008, pp. 641–644.
- [9] K. Babalola, A. Gait, and T. F. Cootes, “A parts-and-geometry initialiser for 3d non-rigid registration using features derived from spin images,” *Neurocomputing*, vol. 120, pp. 113–120, 2013. DOI: 10.1016/j.neucom.2012.08.052.
- [10] W. Bai, W. Shi, A. de Marvao, T. J. W. Dawes, D. P. O’Regan, S. A. Cook, and D. Rueckert, “A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion,” *Medical Image Analysis*, vol. 26, no. 1, pp. 133–145, 2015. DOI: 10.1016/j.media.2015.08.009.
- [11] W. Bai, W. Shi, D. P. O’Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert, “A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac mr images,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1302–1315, 2013. DOI: 10.1109/TMI.2013.2256922.
- [12] R. Bajcsy and S. Kovačič, “Multiresolution elastic matching,” *Computer vision, graphics, and image processing*, vol. 46, no. 1, pp. 1–21, 1989.
- [13] D. Barbosa, D. Friboulet, J. Dhooge, and O. Bernard, “Fast tracking of the left ventricle using global anatomical affine optical flow and local recursive block matching,” *Proceedings of MICCAI CETUS challenge*, 2014.
- [14] D. Barbosa *et al.*, “Fast and fully automatic 3-D echocardiographic segmentation using B-spline explicit active surfaces: Feasibility study and validation in a clinical setting,” *Ultrasound in Medicine & Biology*, vol. 39, no. 1, pp. 89–101, 2013.

- [15] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, “An exploration of 2D and 3D deep learning techniques for cardiac mr image segmentation,” *ArXiv preprint arXiv:1709.04496*, 2017.
- [16] M. F. Beg, M. I. Miller, A. Trouné, and L. Younes, “Computing large deformation metric mappings via geodesic flows of diffeomorphisms,” *International journal of computer vision*, vol. 61, no. 2, pp. 139–157, 2005.
- [17] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, 2003. DOI: 10.1162/089976603321780317.
- [18] N. G. Bellenger, J. M. Francis, C. L. Davies, A. J. Coats, and D. J. Pennell, “Establishment and performance of a magnetic resonance cardiac function clinic,” *Journal of cardiovascular magnetic resonance*, vol. 2, no. 1, pp. 15–22, 2000.
- [19] A. BenTaieb and G. Hamarneh, “Topology aware fully convolutional networks for histology gland segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Springer International Publishing, 2016, pp. 460–468. DOI: 10.1007/978-3-319-46723-8_{_}53.
- [20] O. Bernard, J. Bosch, B. Heyde, M. Alessandrini, D. Barbosa, S. Camarasu-Pop, F. Cervenansky, S. Valette, O. Mirea, M. Bernier, P. Jodoin, J. Domingos, R. Stebbing, K. Keraudren, O. Oktay, J. Caballero, W. Shi, D. Rueckert, F. Milletari, S. Ahmadi, E. Smistad, F. Lindseth, M. van Stralen, C. Wang, O. Smedby, E. Donal, M. Monaghan, A. Papachristidis, M. Geleijnse, E. Galli, and J. Dhooze, “Standardized evaluation system for left ventricular segmentation algorithms in 3d echocardiography,” *IEEE Transactions on Medical Imaging*, vol. 0062, no. c, pp. 1–1, 2015. DOI: 10.1109/TMI.2015.2503890.
- [21] P. B. Bertrand, R. A. Levine, E. M. Isselbacher, and P. M. Vandervoort, “Fact or artifact in two-dimensional echocardiography: Avoiding misdiagnosis and missed diagnosis,” *Journal of the American Society of Echocardiography*, vol. 29, no. 5, pp. 381–391, 2016.

- [22] J. Betancur, A. Simon, F. Schnell, E. Donal, A. Hernández, and M. Garreau, “Evaluation of a motion artifacts removal approach on breath-hold cine-magnetic resonance images of hypertrophic cardiomyopathy subjects,” in *International Seminar on Medical Information Processing and Analysis*, 2013.
- [23] K. K. Bhatia, A. N. Price, W. Shi, and D. Rueckert, “Super-resolution reconstruction of cardiac MRI using coupled dictionary learning,” in *IEEE ISBI*, 2014, pp. 947–50.
- [24] A. Bistoquet, J. Oshinski, and O. Skrinjar, “Left ventricular deformation recovery from cine MRI using an incompressible model,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 9, pp. 1136–1153, 2007.
- [25] Å. Björck, *Numerical methods for least squares problems*. SIAM, 1996.
- [26] F. L. Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 6, pp. 567–585, 1989.
- [27] W. F. Boron and E. L. Boulpaep, *Medical Physiology, 3rd Edition*. Elsevier Health Sciences, 2016.
- [28] Y. Y. Boykov and M.-P. Jolly, “Interactive graph cuts for optimal boundary & region segmentation of objects in nd images,” in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, IEEE, vol. 1, 2001, pp. 105–112.
- [29] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [30] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [31] L. Breimann, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.
- [32] J. Bromley, I. Guyon, Y. Lecun, E. Sckinger, and R. Shah, “Signature verification using a Siamesei time delay neural network,” *NIPS*, pp. 737–744, 1994.

- [33] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan, *Magnetic resonance imaging: Physical principles and sequence design*. John Wiley & Sons, 2014.
- [34] A. Buades, B. Coll, and J.-M. Morel, “A non-local algorithm for image denoising,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 60–65.
- [35] J. Caballero, W. Bai, A. N. Price, D. Rueckert, and J. V. Hajnal, “Application-driven mri: Joint reconstruction and segmentation from undersampled mri data,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2014, pp. 106–113.
- [36] P. Cachier and X. Pennec, “3D non-rigid registration by gradient descent on a Gaussian windowed similarity measure using convolutions,” in *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2000, pp. 182–89.
- [37] M. C. Carminati, F. Maffessanti, and E. G. Caiani, “Automated motion artifacts removal between cardiac long- and short-axis magnetic resonance images,” *Computing in Cardiology*, pp. 689–692, 2012.
- [38] 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*, vol. 219, no. 3, pp. 828–834, 2001.
- [39] E. Castillo and D. A. Bluemke, “Cardiac MR imaging,” *Radiologic Clinics*, vol. 41, no. 1, pp. 17–28, 2003.
- [40] E. Castillo, J. A. Lima, and D. A. Bluemke, “Regional myocardial function: Advances in MR imaging and analysis,” *Radiographics*, vol. 23, no. suppl.1, S127–S140, 2003.
- [41] M. D. Cerqueira, N. J. Weissman, V. Dilsizian, A. K. Jacobs, S. Kaul, W. K. Laskey, D. J. Pennell, J. A. Rumberger, T. Ryan, and M. S. Verani, “Standardized myocardial segmentation and nomenclature for tomographic imaging of the heart,” *Circulation*, vol. 105, no. 4, 2002.

- [42] A. G. Chandler, R. J. Pinder, T. Netsch, J. A. Schnabel, D. J. Hawkes, D. L. G. Hill, and R. Razavi, "Correction of misaligned slices in multi-slice cardiovascular magnetic resonance using slice-to-volume registration.," *Journal of cardiovascular magnetic resonance*, vol. 10, p. 13, 2008.
- [43] F. Chen, H. Yu, R. Hu, and X. Zeng, "Deep learning shape priors for object segmentation," in *IEEE CVPR*, 2013, pp. 1870–77.
- [44] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. A. Heng, "Dcan: deep contour-aware networks for object instance segmentation from histology images," *Medical Image Analysis*, vol. 36, pp. 135–146, 2017. DOI: 10.1016/j.media.2016.11.004. arXiv: 1604.02677.
- [45] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, and S. K. Zhou, "Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*. Springer International Publishing, 2016, pp. 487–495. DOI: 10.1007/978-3-319-46723-8{_}56.
- [46] T. Chen, J. Babb, P. Kellman, L. Axel, and D. Kim, "Semiautomated segmentation of myocardial contours for fast strain analysis in cine displacement-encoded mri," *IEEE Transactions on Medical Imaging*, vol. 27, no. 8, pp. 1084–1094, 2008.
- [47] E. Chou and J. Carrino, *Magnetic Resonance Imaging*. Elsevier Inc., 2006, vol. 1, pp. 106–117.
- [48] G. E. Christensen, R. D. Rabbitt, and M. I. Miller, "Deformable templates using large deformation kinematics," *IEEE Transactions on Image Processing*, vol. 5, no. 10, pp. 1435–1447, 1996.
- [49] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Confer-*

- ence, Athens, Greece, October 17-21, 2016, *Proceedings, Part II*. Springer International Publishing, 2016, pp. 424–432. DOI: 10.1007/978-3-319-46723-8{_}49.
- [50] S. Conjeti, A. Katouzian, A. Kazi, S. Mesbah, D. Beymer, T. F. Syeda-Mahmood, and N. Navab, “Metric hashing forests,” *Medical Image Analysis*, vol. 34, pp. 13–29, 2016.
- [51] G. Constantine, K. Shan, S. D. Flamm, and M. U. Sivananthan, “Role of MRI in clinical cardiology,” *The Lancet*, vol. 363, no. 9427, pp. 2162–2171, 2004.
- [52] T. F. Cootes and C. J. Taylor, “Combining point distribution models with shape models based on finite element analysis,” *Image and Vision Computing*, vol. 13, no. 5, pp. 403–409, 1995. DOI: 10.1016/0262-8856(95)99727-I.
- [53] T. F. Cootes, C. J. Twining, V. S. Petrovic, K. O. Babalola, and C. J. Taylor, “Computing accurate correspondences across groups of images,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1994–2005, 2010. DOI: 10.1109/TPAMI.2009.193.
- [54] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, “Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation,” *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011. DOI: 10.1016/j.neuroimage.2010.09.018.
- [55] A. M. Crean, N. Maredia, G. Ballard, R. Menezes, G. Wharton, J. Forster, J. P. Greenwood, and J. D. Thomson, “3D echo systematically underestimates right ventricular volumes compared to cardiovascular magnetic resonance in adult congenital heart disease patients with moderate or severe RV dilatation,” *Journal of Cardiovascular Magnetic Resonance*, vol. 13, no. 1, p. 78, 2011.
- [56] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, “Regression forests for efficient anatomy detection and localization in computed tomography scans,” *Medical Image Analysis*, vol. 17, no. 8, pp. 1293–1303, 2013. DOI: 10.1016/j.media.2013.01.001.
- [57] A. Criminisi, *Machine learning for medical images analysis*, 2016.

- [58] A. Criminisi and J. Shotton, *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media, 2013.
- [59] A. Criminisi, J. Shotton, and S. Bucciarelli, “Decision forests with long-range spatial context for organ localization in ct volumes,” in *MICCAI workshop on Probabilistic Models for Medical Image Analysis (MICCAI-PMMIA)*, 2009.
- [60] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1, IEEE, 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.
- [61] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, “Real-time facial feature detection using conditional regression forests,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 2578–2585. DOI: 10.1109/CVPR.2012.6247976.
- [62] C. Davatzikos, X. Tao, and D. Shen, “Hierarchical active shape models, using the wavelet transform,” *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 414–423, 2003. DOI: 10.1109/TMI.2003.809688.
- [63] D. De Nigris, D. L. Collins, and T. Arbel, “Multi-modal image registration based on gradient orientations of minimal uncertainty,” *IEEE TMI*, vol. 31, no. 12, pp. 2343–54, 2012.
- [64] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.
- [65] T. Dietenbeck, M. Alessandrini, D. Barbosa, J. Dhooge, D. Friboulet, and O. Bernard, “Detection of the whole myocardium in 2d-echocardiography for multiple orientations using a geometrically constrained level-set,” *Medical Image Analysis*, vol. 16, no. 2, pp. 386–401, 2012.
- [66] P. Dollar, Z. Tu, P. Perona, and S. Belongie, “Integral channel features,” in *Proceedings of the British Machine Vision Conference 2009*, British Machine Vision Association, 2009, pp. 91.1–91.11. DOI: 10.5244/C.23.91.

- [67] P. Dollar and C. L. Zitnick, “Structured forests for fast edge detection,” in *The IEEE International Conference on Computer Vision (ICCV)*, 2013.
- [68] J. S. Domingos, R. V. Stebbing, P. Leeson, and J. A. Noble, “Structured random forests for myocardium delineation in 3D echocardiography,” in *MLMI*, 2014.
- [69] C. Dong, Y. Deng, C. Change Loy, and X. Tang, “Compression artifacts reduction by a deep convolutional network,” in *IEEE CVPR*, 2015, pp. 576–584.
- [70] C. Dong, C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE PAMI*, vol. 38, no. 2, pp. 295–307, 2016.
- [71] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in *European Conference on Computer Vision*, Springer, 2014, pp. 184–199.
- [72] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II*. Springer International Publishing, 2016, pp. 391–407. DOI: 10.1007/978-3-319-46475-6{_}25.
- [73] R. Donner, B. H. Menze, H. Bischof, and G. Langs, “Global localization of 3d anatomical structures by pre-filtered hough forests and discrete optimization,” *Medical Image Analysis*, vol. 17, no. 8, pp. 1304–1314, 2013. DOI: 10.1016/j.media.2013.02.004.
- [74] I. Dydenko, F. Jamal, O. Bernard, J. Dhooge, I. E. Magnin, and D. Friboulet, “A level set framework with a shape and motion prior for segmentation and region tracking in echocardiography,” *Medical image analysis*, vol. 10, no. 2, pp. 162–177, 2006.
- [75] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–45, 2006.
- [76] S Eslami and C. Williams, “A generative model for parts-based object segmentation,” in *NIPS*, 2012, pp. 100–107.

- [77] S. M. A. Eslami, N. Heess, and J. Winn, “The shape boltzmann machine: A strong model of object shape,” in *IEEE CVPR*, 2012, pp. 406–413.
- [78] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*.
- [79] E. Ferrante and N. Paragios, “Slice-to-volume medical image registration: A survey,” *Medical Image Analysis*, vol. 39, pp. 101–123, 2017.
- [80] J. P. Finn, K. Nael, V. Deshpande, O. Ratib, and G. Laub, “Cardiac MR imaging: State of the technology,” *Radiology*, vol. 241, no. 2, pp. 338–354, 2006.
- [81] J. M. Fitzpatrick, J. B. West, and C. R. Maurer, “Predicting error in rigid-body point-based registration,” *IEEE transactions on medical imaging*, vol. 17, no. 5, pp. 694–702, 1998.
- [82] J. Folkesson, E. Samset, R. Y. Kwong, and C.-F. Westin, “Unifying statistical classification and geodesic active regions for segmentation of cardiac mri,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 12, no. 3, pp. 328–334, 2008.
- [83] 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, D. C. Lee, J. A. C. Lima, P. Medrano-Gracia, K. Shivkumar, A. Suinesiaputra, W. Tao, and A. A. Young, “The cardiac atlas project-an imaging database for computational modeling and statistical atlases of the heart,” *Bioinformatics*, vol. 27, no. 16, pp. 2288–2295, 2011. DOI: 10.1093/bioinformatics/btr360.
- [84] M. Foppa, B. B. Duncan, and L. E. Rohde, “Echocardiography-based left ventricular mass estimation. how should we define hypertrophy?” *Cardiovascular ultrasound*, vol. 3, no. 1, p. 17, 2005.
- [85] 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*, vol. 20, no. 1, pp. 2–5, 2001.

- [86] A. F. Frangi, D. Rueckert, J. A. Schnabel, and W. J. Niessen, “Automatic construction of multiple-object three-dimensional statistical shape models: Application to cardiac modeling,” *IEEE transactions on medical imaging*, vol. 21, no. 9, pp. 1151–1166, 2002.
- [87] B. Fuerst, W. Wein, M. Müller, and N. Navab, “Automatic ultrasound - MRI registration for neurosurgery using the 2D and 3D LC² metric,” *MedIA*, 2014.
- [88] K. Fukushima and S. Miyake, “Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition,” in *Competition and cooperation in neural nets*, Springer, 1982, pp. 267–285.
- [89] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, 2016.
- [90] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, “Hough forests for object detection, tracking, and action recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011. DOI: 10.1109/TPAMI.2011.70.
- [91] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, “Multi-organ localization combining global-to-local regression and confidence maps,” *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings*, pp. 337–344, 2014.
- [92] F. C. Ghesu, B. Georgescu, T. Mansi, D. Neumann, J. Hornegger, and D. Comaniciu, “An artificial agent for anatomical landmark detection in medical images,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp. 229–237.
- [93] 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*, vol. 197, no. 4, pp. 808–815, 2011.

- [94] R. Girdhar, D. F. Fouhey, M. Rodriguez, and A. Gupta, “Learning a predictable and generative vector representation for objects,” in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*. Springer International Publishing, 2016, pp. 484–499. DOI: 10.1007/978-3-319-46466-4{_}29.
- [95] R. Girshick, A. Criminisi, J. Shotton, A. Fitzgibbon, and P. Kohli, “Efficient regression of general-activity human poses from depth images,” *2011 IEEE International Conference on Computer Vision (ICCV 2011)*, vol. 00, pp. 415–422, 2011. DOI: doi.ieeecomputersociety.org/10.1109/ICCV.2011.6126270.
- [96] A. C. Glatz, X. Zhu, M. J. Gillespie, B. D. Hanna, and J. J. Rome, “Use of angiographic ct imaging in the cardiac catheterization laboratory for congenital heart disease,” *JACC: Cardiovascular Imaging*, vol. 3, no. 11, pp. 1149–1157, 2010.
- [97] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi, “Joint classification-regression forests for spatially structured multi-object segmentation,” in *Computer Vision – ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV*. Springer Berlin Heidelberg, 2012, pp. 870–881. DOI: 10.1007/978-3-642-33765-9{_}62.
- [98] S. Gold, C.-P. Lu, A. Rangarajan, S. Pappu, and E. Mjolsness, “New algorithms for 2d and 3d point matching: Pose estimation and correspondence,” in *Advances in Neural Information Processing Systems*, 1995, pp. 957–964.
- [99] A. Gomez, O. Oktay, D. Rueckert, G. P. Penney, J. A. Schnabel, J. M. Simpson, and K. Pushparajah, “Regional differences in end-diastolic volumes between 3d echo and cmr in hlhs patients,” *Frontiers in Pediatrics*, vol. 4, 2016.
- [100] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.

- [101] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
- [102] J. C. Gower, “Generalized procrustes analysis,” *Psychometrika*, vol. 40, no. 1, pp. 33–51, 1975.
- [103] H. Greenspan, “Super-resolution in medical imaging,” *The Computer Journal*, vol. 52, no. 1, pp. 43–63, 2008. DOI: 10.1093/comjnl/bxm075.
- [104] 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*, vol. 35, no. 5, pp. 1153–1159, 2016.
- [105] H. Greenspan, G. Oz, N. Kiryati, and S. Peled, “MRI inter-slice reconstruction using super-resolution,” *Magnetic resonance imaging*, vol. 20, no. 5, pp. 437–446, 2002.
- [106] H. Heinrich, M. Jenkinson, M. Brady, and J. A. Schnabel, “MRF-based deformable registration and ventilation estimation of lung CT,” *IEEE TMI*, pp. 1239–48, 2013.
- [107] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel, “Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration,” *Medical Image Analysis*, vol. 16, no. 7, pp. 1423–1435, 2012.
- [108] M. P. Heinrich and O. Oktay, “Briefnet: Deep pancreas segmentation using binary sparse convolutions,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 329–337.
- [109] M. P. Heinrich, M. Jenkinson, B. W. Papież, M. Brady, and J. A. Schnabel, “Towards realtime multimodal fusion for image-guided interventions using self-similarities,” in *MICCAI*, 2013, pp. 187–94.
- [110] P. Hellier, C. Barillot, É. Mémin, and P. Pérez, “Hierarchical estimation of a dense deformation field for 3-d robust registration,” *IEEE transactions on medical imaging*, vol. 20, no. 5, pp. 388–402, 2001.

- [111] K. Hergan, A. Schuster, J. Frühwald, M. Mair, R. Burger, and M. Töpker, “Comparison of left and right ventricular volume measurement using the simpson’s method and the area length method,” *European journal of radiology*, vol. 65, no. 2, pp. 270–278, 2008.
- [112] C. B. Higgins and A. de Roos, *MRI and CT of the Cardiovascular System*. Lippincott Williams & Wilkins, 2006.
- [113] T. K. Ho, “Random decision forests,” in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, IEEE, vol. 1, 1995, pp. 278–282.
- [114] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [115] P. R. Hoskins, K. Martin, and A. Thrush, *Diagnostic ultrasound: Physics and equipment*. Cambridge University Press, 2010.
- [116] R. J. Housden, M. Basra, Y. Ma, A. P. King, R. Bullens, N. Child, J. Gill, C. A. Rinaldi, V. Parish, and K. S. Rhode, “Three-modality registration for guidance of minimally invasive cardiac interventions,” in *International Conference on Functional Imaging and Modeling of the Heart*, Springer, 2013, pp. 158–165.
- [117] D. H. Hubel and T. N. Wiesel, “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex,” *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.
- [118] F. Isensee, P. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, “Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features,” *ArXiv preprint arXiv:1707.00587*, 2017.
- [119] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. Van Ginneken, “Multi-atlas-based segmentation with local decision fusion application to cardiac and aortic segmentation in ct scans,” *IEEE transactions on medical imaging*, vol. 28, no. 7, pp. 1000–1010, 2009.

- [120] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*. Springer International Publishing, 2016, pp. 694–711. DOI: 10.1007/978-3-319-46475-6{_}43.
- [121] M.-P. Jolly, “Automatic segmentation of the left ventricle in cardiac mr and ct images,” *International Journal of Computer Vision*, vol. 70, no. 2, pp. 151–163, 2006.
- [122] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, *et al.*, “Unsupervised domain adaptation in brain lesion segmentation with adversarial networks,” in *International Conference on Information Processing in Medical Imaging*, Springer, 2017, pp. 597–609.
- [123] K. Kamnitsas, C. Ledig, V. F. 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*, vol. 36, pp. 61–78, 2017. DOI: 10.1016/j.media.2016.10.004. arXiv: 1603.05959.
- [124] T. D. Karamitsos, J. M. Francis, S. Myerson, J. B. Selvanayagam, and S. Neubauer, *The role of cardiovascular magnetic resonance imaging in heart failure*, 2009. DOI: 10.1016/j.jacc.2009.04.094.
- [125] A. Kendall, Y. Gal, and R. Cipolla, “Multi-task learning using uncertainty to weigh losses for scene geometry and semantics,” *ArXiv preprint arXiv:1705.07115*, 2017.
- [126] J. A. Kennedy, O. Israel, A. Frenkel, R. Bar-Shalom, and H. Azhari, “Super-resolution in PET imaging,” *IEEE Transactions on Medical Imaging*, vol. 25, no. 2, pp. 137–147, 2006.
- [127] A. P. King, K. S. Rhode, Y. Ma, C. Yao, C. Jansen, R. Razavi, and G. P. Penney, “Registering preprocedure volumetric images with intraprocedure 3-d ultrasound using an ultrasound imaging model,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 3, pp. 924–937, 2010.

- [128] S. Klein, U. A. van der Heide, I. M. Lips, M. van Vulpen, M. Staring, and J. P. Pluim, “Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information,” *Medical Physics*, vol. 35, no. 4, pp. 1407–17, 2008.
- [129] P. Kotschieder, S. Rota Bulo, H. Bischof, and M. Pelillo, “Structured class-labels in random forests for semantic image labelling,” in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 2190–2197. DOI: 10.1109/ICCV.2011.6126496.
- [130] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi, “Neighbourhood approximation using randomized forests,” *Medical Image Analysis*, vol. 17, no. 7, pp. 790–804, 2013. DOI: 10.1016/j.media.2013.04.013.
- [131] P. Kovesi *et al.*, “Symmetry and asymmetry from local phase,” in *Tenth Australian joint conference on artificial intelligence*, Citeseer, vol. 190, 1997, pp. 2–4.
- [132] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., 2012, pp. 1097–1105.
- [133] M. Kuklisova-Murgasova, A. Cifor, R. Napolitano, A. Papageorgiou, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, J. A. Noble, and J. A. Schnabel, “Registration of 3d fetal neurosonography and mri,” *Medical image analysis*, vol. 17, no. 8, pp. 1137–1150, 2013.
- [134] M. Kuklisova-Murgasova, G. Quaghebeur, M. A. Rutherford, J. V. Hajnal, and J. A. Schnabel, “Reconstruction of fetal brain mri with intensity matching and complete outlier removal,” *Medical image analysis*, vol. 16, no. 8, pp. 1550–1564, 2012.
- [135] J. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional random fields: probabilistic models for segmenting and labeling sequence data,” *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, vol. 8, no. June, pp. 282–289, 2001. DOI: 10.1038/nprot.2006.61. arXiv: arXiv:1011.4088v1.

- [136] P. Lancellotti, C. Tribouilloy, A. Hagendorff, B. A. Popescu, T. Edvardsen, L. A. Pierard, L. Badano, and J. L. Zamorano, “Recommendations for the echocardiographic assessment of native valvular regurgitation: An executive summary from the european association of cardiovascular imaging,” *European Heart Journal–Cardiovascular Imaging*, vol. 14, no. 7, pp. 611–644, 2013.
- [137] R. M. Lang, L. P. Badano, W. Tsang, D. H. Adams, E. Agricola, T. Buck, F. F. Faletra, A. Franke, J. Hung, L. Pérez De Isla, O. Kamp, J. D. Kasprzak, P. Lancellotti, T. H. Marwick, M. L. McCulloch, M. J. Monaghan, P. Nihoyannopoulos, N. G. Pandian, P. A. Pellikka, M. Pepi, D. A. Roberson, S. K. Shernan, G. S. Shirali, L. Sugeng, F. J. Ten Cate, M. A. Vannan, J. L. Zamorano, and W. A. Zoghbi, “Eae/ase recommendations for image acquisition and display using three-dimensional echocardiography,” *Journal of the American Society of Echocardiography*, vol. 25, no. 1, pp. 3–46, 2012. DOI: 10.1016/j.echo.2011.11.010.
- [138] M. Le, J. Lieman-Sifry, F. Lau, S. Sall, A. Hsiao, and D. Golden, “Computationally efficient cardiac views projection using 3d convolutional neural networks,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer, 2017, pp. 109–116.
- [139] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [140] C. Ledig, W. Shi, W. Bai, and D. Rueckert, “Patch-based evaluation of image segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3065–3072.
- [141] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” *CoRR*, vol. abs/1609.0, 2016.
- [142] H.-Y. Lee, N. C. Codella, M. D. Cham, J. W. Weinsaft, and Y. Wang, “Automatic left ventricle segmentation using iterative thresholding and an active contour model with

- adaptation on short-axis cardiac mri,” *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 905–913, 2010.
- [143] K. Lekadir, R. Merrifield, and G.-Z. Yang, “Outlier detection and handling for robust 3-d active shape models search,” *IEEE Transactions on Medical Imaging*, vol. 26, no. 2, pp. 212–222, 2007.
- [144] V. Lempitsky, M. Verhoek, J. A. Noble, and A. Blake, “Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography,” in *FIMH*, 2009, pp. 447–56.
- [145] M. Leordeanu and M. Hebert, “A spectral technique for correspondence problems using pairwise constraints,” in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, IEEE, vol. 2, 2005, pp. 1482–1489.
- [146] K. Li, B. Hariharan, and J. Malik, “Iterative instance segmentation,” in *CVPR*, 2016, pp. 3659–3667.
- [147] S. Z. Li, “Markov random field models in computer vision,” in *European conference on computer vision*, Springer, 1994, pp. 361–370.
- [148] Y. Li and R. Verma, “Multichannel image registration by feature-based information fusion,” *IEEE Transactions on Medical Imaging*, vol. 30, no. 3, pp. 707–720, 2011.
- [149] Y. Li, M. Paluri, J. M. Rehg, and P. Dollar, “Unsupervised learning of edges,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [150] G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *CoRR*, vol. abs/1702.0, 2017.
- [151] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Suetens, “Nonrigid image registration using conditional mutual information,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 19–29, 2010.

- [152] H. Lombaert, Y. Sun, and F. Cheriet, “Fast 4d segmentation of large datasets using graph cuts,” in *Medical Imaging: Image Processing*, 2011, 79622H.
- [153] H. Lombaert, D. Zikic, A. Criminisi, and N. Ayache, “Laplacian forests: semantic image segmentation by guided bagging,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II*. Springer International Publishing, 2014, pp. 496–504. DOI: 10.1007/978-3-319-10470-6{_}62.
- [154] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [155] M. Lorenzo-Valdés, G. I. Sanchez-Ortiz, A. G. Elkington, R. H. Mohiaddin, and D. Rueckert, “Segmentation of 4d cardiac mr images using a probabilistic atlas and the em algorithm,” *Medical Image Analysis*, vol. 8, no. 3, pp. 255–265, 2004.
- [156] J. Lötjönen, M. Pollari, and K. Lauerma, “Correction of movement artifacts from 4-D cardiac short-and long-axis MR data,” in *MICCAI*, Springer, 2004, pp. 405–12.
- [157] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [158] A. Lu, M. Zontak, N. Parajuli, J. C. Stendahl, N. Boutagy, M. Eberle, I. Alkhalil, M. ODonnell, A. J. Sinusas, and J. S. Duncan, “Learning-based spatiotemporal regularization and integration of tracking methods for regional 4d cardiac deformation analysis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 323–331.
- [159] J.-y. Lu, H. Zou, and J. F. Greenleaf, “Biomedical ultrasound beam forming,” *Ultrasound in medicine & biology*, vol. 20, no. 5, pp. 403–428, 1994.
- [160] X. Lu, B. Georgescu, M. P. Jolly, J. Guehring, A. Young, B. Cowan, A. Littmann, and D. Comaniciu, “Cardiac anchoring in MRI through context modeling,” in *Lecture Notes in Computer Science*, vol. 6361 LNCS, 2010, pp. 383–390.

- [161] X. Lu and M. P. Jolly, “Discriminative context modeling using auxiliary markers for lv landmark detection from a single mr image,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7746 LNCS, 2013, pp. 105–114. DOI: 10.1007/978-3-642-36961-2{_}13.
- [162] X. Lu, M.-p. Jolly, B. Georgescu, C. Hayes, P. Speier, M. Schmidt, X. Bi, R. Kroeker, D. Comaniciu, P. Kellman, E. Mueller, and J. Guehring, “Automatic view planning for cardiac mri acquisition,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011: 14th International Conference*, 2011, pp. 479–486.
- [163] Y.-L. Lu, K. A. Connelly, A. J. Dick, G. A. Wright, and P. E. Radau, “Automatic functional analysis of left ventricle in cardiac cine mri,” *Quantitative imaging in medicine and surgery*, vol. 3, no. 4, p. 200, 2013.
- [164] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, “Semantic segmentation using adversarial networks,” *CoRR*, vol. abs/1611.0, 2016.
- [165] B. D. Lucas, T. Kanade, *et al.*, “An iterative image registration technique with an application to stereo vision,” 1981.
- [166] M. Lustig, D. Donoho, and J. M. Pauly, “Sparse MRI: The application of compressed sensing for rapid MR imaging,” *Magnetic resonance in medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.
- [167] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, “Multimodality image registration by maximization of mutual information,” *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.
- [168] D. Mahapatra, “Cardiac image segmentation from cine cardiac mri using graph cuts and shape priors,” *Journal of digital imaging*, vol. 26, no. 4, pp. 721–730, 2013.
- [169] D. Mahapatra and Y. Sun, “Orientation histograms as shape priors for left ventricle segmentation using graph cuts,” *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*, pp. 420–427, 2011.

- [170] S. Malm, S. Frigstad, E. Sagberg, H. Larsson, and T. Skjaerpe, “Accurate and reproducible measurement of left ventricular volume and ejection fraction by contrast echocardiography,” *Journal of the American College of Cardiology*, vol. 44, no. 5, pp. 1030–1035, 2004.
- [171] J. V. Manjón, P. Coupé, A. Buades, V. Fonov, D. L. Collins, and M. Robles, “Non-local MRI upsampling,” *MedIA*, vol. 14, no. 6, pp. 784–792, 2010.
- [172] S. Marchesseau, J. X. Ho, and J. J. Totman, “Influence of the short-axis cine acquisition protocol on the cardiac function evaluation: A reproducibility study,” *European journal of radiology open*, vol. 3, pp. 60–66, 2016.
- [173] J. Margeta, E. Geremia, A. Criminisi, and N. Ayache, “Layered spatio-temporal forests for left ventricle segmentation from 4D cardiac MRI data,” in *STACOM, Imaging and Modelling Challenges*, 2012, pp. 109–19.
- [174] A. de Marvao, T. J. Dawes, W. Shi, C. Minas, N. G. Keenan, T. Diamond, G. Durighel, G. Montana, D. Rueckert, S. A. Cook, and D. P. ORegan, “Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power,” *Journal of Cardiovascular Magnetic Resonance*, vol. 16, no. 1, p. 16, 2014. DOI: 10.1186/1532-429X-16-16.
- [175] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” in *Artificial Neural Networks and Machine Learning – ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I*. Springer Berlin Heidelberg, 2011, pp. 52–59. DOI: 10.1007/978-3-642-21735-7{_}7.
- [176] B. H. Menze, B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht, “On oblique random forests,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2011, pp. 453–469.
- [177] F. Milletari, S.-A. Ahmadi, C. Kroll, C. Hennemersperger, F. Tombari, A. Shah, A. Plate, K. Boetzel, and N. Navab, “Robust segmentation of various anatomies in 3d ultrasound

- using hough forests and learned data representations,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 111–118.
- [178] F. Milletari, A. Rothberg, J. Jia, and M. Sofka, “Integrating statistical prior knowledge into convolutional neural networks,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 161–168.
- [179] Min Sun, P. Kohli, and J. Shotton, “Conditional regression forests for human pose estimation,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2012, pp. 3394–3401. DOI: 10.1109/CVPR.2012.6248079.
- [180] S. C. Mitchell, J. G. Bosch, B. P. Lelieveldt, R. J. Van der Geest, J. H. Reiber, and M. Sonka, “3-d active appearance models: Segmentation of cardiac mr and ultrasound images,” *IEEE transactions on medical imaging*, vol. 21, no. 9, pp. 1167–1178, 2002.
- [181] A. Myronenko and X. Song, “Point set registration: Coherent point drift,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.
- [182] J. C. Nascimento and G. Carneiro, “Multi-atlas segmentation using manifold learning with deep belief networks,” in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, IEEE, 2016, pp. 867–871.
- [183] T. R. Nelson, D. H. Pretorius, A. Hull, M. Riccabona, M. S. Sklansky, and G. James, “Sources and impact of artifacts on clinical three-dimensional ultrasound imaging,” *Ultrasound in Obstetrics and Gynecology*, vol. 16, no. 4, pp. 374–383, 2000. DOI: 10.1046/j.1469-0705.2000.00180.x.
- [184] C. T. Nguyen, E. Lee, H. Luo, and R. J. Siegel, “Echocardiographic guidance for diagnostic and therapeutic percutaneous procedures,” *Cardiovasc Diagn Ther*, 2011.
- [185] J. A. Noble and D. Boukerroui, “Ultrasound image segmentation: A survey,” *IEEE Transactions on medical imaging*, vol. 25, no. 8, pp. 987–1010, 2006.

- [186] J. A. Noble, N. Navab, and H. Becher, “Ultrasonic image analysis and image-guided interventions,” *Interface focus*, vol. 1, no. 4, pp. 673–685, 2011.
- [187] M. S. Nosrati and G. Hamarneh, “Incorporating prior knowledge in medical image segmentation: a survey,” *CoRR*, vol. abs/1607.0, 2016.
- [188] F. Odille, A. Bustin, B. Chen, P.-A. Vuissoz, and J. Felblinger, “Motion-corrected, super-resolution reconstruction for high-resolution 3D cardiac cine MRI,” in *MICCAI*, Springer, 2015, pp. 435–442.
- [189] O. Oktay, A. Schuh, M. Rajchl, K. Keraudren, A. Gomez, M. Heinrich, G. Penney, and D. Rueckert, *Structured decision forests for multi-modal ultrasound image registration*. 2015, vol. 9350.
- [190] O. Oktay, W. Bai, R. Guerrero, M. Rajchl, A. De Marvao, D. P. O’Regan, S. A. Cook, M. P. Heinrich, B. Glocker, and D. Rueckert, “Stratified decision forests for accurate anatomical landmark localization in cardiac images,” *IEEE Transactions on Medical Imaging*, vol. 36, no. 1, pp. 332–342, 2017. DOI: 10.1109/TMI.2016.2597270.
- [191] O. Oktay, W. Bai, M. Lee, R. Guerrero, K. Kamnitsas, J. Caballero, A. de Marvao, S. Cook, D. O’Regan, and D. Rueckert, “Multi-input cardiac image super-resolution using convolutional neural networks,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III*. Springer International Publishing, 2016, pp. 246–254. DOI: 10.1007/978-3-319-46726-9{_}29.
- [192] O. Oktay, A. Gomez, K. Keraudren, A. Schuh, W. Bai, W. Shi, G. Penney, and D. Rueckert, “Probabilistic edge map (pem) for 3d ultrasound image registration and multi-atlas left ventricle segmentation,” in *International Conference on Functional Imaging and Modeling of the Heart*, Springer, 2015, pp. 223–230.
- [193] O. Oktay, W. Shi, J. Caballero, K. Keraudren, and D. Rueckert, “Sparsity based spectral embedding: Application to multi-atlas echocardiography segmentation,” in *Proceedings of MICCAI STMI Workshop*, 2014.

- [194] L. H. Opie, *Heart physiology: From cell to circulation*. Lippincott Williams & Wilkins, 2004.
- [195] A Oppelt, R Graumann, H Barfuss, H Fischer, W Hartl, W Schajor, *et al.*, “FISPa new fast MRI sequence,” *Electromedica*, vol. 54, no. 1, pp. 15–18, 1986.
- [196] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos, “Dramms: Deformable registration via attribute matching and mutual-saliency weighting,” *Medical image analysis*, vol. 15, no. 4, pp. 622–639, 2011.
- [197] S. Ourselin, A. Roche, S. Prima, and N. Ayache, “Block matching: a general framework to improve robustness of rigid registration of medical images,” *MICCAI 2000*, vol. 1935, pp. 557–566, 2000. arXiv: [arXiv:1011.1669v3](https://arxiv.org/abs/1011.1669v3).
- [198] S. Ourselin, A. Roche, X. Pennec, and N. Ayache, “Reconstructing a 3D structure from serial histological sections,” *Image and Vision Computing*, vol. 19, no. 1, pp. 25–31, 2001.
- [199] A. Papachristidis *et al.*, “Clinical expert delineation of 3D left ventricular echocardiograms for the CETUS segmentation challenge,” *Proceedings of MICCAI CETUS challenge*, pp. 9–16, 2014.
- [200] C. H. Papadimitriou and K. Steiglitz, “The max-flow, min-cut theorem,” in *Combinatorial Optimization: Algorithms and Complexity*, Courier Corporation, 1998, ch. 6, pp. 120–128.
- [201] T. Papavassiliu, H. P. Kuhl, M. Schroder, T. Suselbeck, O. Bondarenko, C. K. Bohm, A. Beek, M. M. Hofman, and A. C. van Rossum, “Effect of endocardial trabeculae on left ventricular measurements and measurement reproducibility at cardiovascular mr imaging,” *Radiology*, vol. 236, no. 1, pp. 57–64, 2005.
- [202] N. Paragios, “A variational approach for the segmentation of the left ventricle in cardiac image analysis,” *International Journal of Computer Vision*, vol. 50, no. 3, pp. 345–362, 2002.

- [203] N. Paragios, M.-P. Jolly, M. Taron, and R. Ramaraj, "Active shape models and segmentation of the left ventricle in echocardiography," in *Scale-Space*, Springer, 2005, pp. 131–142.
- [204] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.
- [205] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 155–195, 2016.
- [206] 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, R. Collins, S. Piechnik, and S. Neubauer, "Uk biobanks cardiovascular magnetic resonance protocol," *Journal of Cardiovascular Magnetic Resonance*, vol. 18, no. 1, p. 8, 2015. DOI: 10.1186/s12968-016-0227-4.
- [207] C. Petitjean, M. A. Zuluaga, W. Bai, J.-N. Dacher, D. Grosgeorge, J. Caudron, S. Ruan, I. B. Ayed, M. J. Cardoso, H.-C. Chen, *et al.*, "Right ventricle segmentation from cardiac mri: A collation study," *Medical image analysis*, vol. 19, no. 1, pp. 187–202, 2015.
- [208] E. Plenge, D. Poot, W. Niessen, and E. Meijering, "Super-resolution reconstruction using cross-scale self-similarity in multi-slice MRI," in *MICCAI*, 2013, pp. 123–130.
- [209] V. Potesil, T. Kadir, G. Platsch, and M. Brady, "Personalized graphical models for anatomical landmark localization in whole-body medical images," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 29–49, 2014. DOI: 10.1007/s11263-014-0731-7.
- [210] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *International Workshop on Reconstruction and Analysis of Moving Body Organs*, Springer, 2016, pp. 83–94.
- [211] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, P. Boesiger, *et al.*, "Sense: Sensitivity encoding for fast MRI," *Magnetic resonance in medicine*, vol. 42, no. 5, pp. 952–962, 1999.

- [212] S. Pszczolkowski, L. Pizarro, R. Guerrero, and D. Rueckert, “Nonrigid free-form registration using landmark-based statistical deformation models,” vol. 8314, 2012, pp. 8314–9.
- [213] E. Puyol-Antón, M. Sinclair, B. Gerber, M. S. Amzulescu, H. Langet, M. De Craene, P. Aljabar, P. Piro, and A. P. King, “A multimodal spatiotemporal cardiac motion atlas from mr and ultrasound data,” *Medical Image Analysis*, 2017.
- [214] T. Rainforth and F. Wood, “Canonical correlation forests,” *ArXiv preprint arXiv:1507.05444*, 2015.
- [215] M. Rajchl, J. Yuan, and T. M. Peters, “Real-time segmentation in 4D ultrasound with continuous max-flow,” in *Medical Imaging: Image Processing*, 2012, 83141F.
- [216] K. Rajpoot, V. Grau, J. Alison Noble, H. Becher, and C. Szmigielski, “The evaluation of single-view and multi-view fusion 3D echocardiography using image-driven segmentation and tracking,” *Medical Image Analysis*, vol. 15, no. 4, pp. 514–28, 2011.
- [217] H. Ravishankar, R. Venkataramani, S. Thiruvankadam, P. Sudhakar, and V. Vaidya, “Learning and incorporating shape models for semantic segmentation,” *ResearchGate*, 2017.
- [218] H. Ravishankar, S. Thiruvankadam, R. Venkataramani, and V. Vaidya, “Joint deep learning of foreground, background and shape for robust contextual segmentation,” in *IPMI*, 2017, pp. 622–632.
- [219] 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*, vol. 112, no. 6, pp. 855–861, 2005.
- [220] J. P. Ridgway, “Cardiovascular magnetic resonance physics for clinicians: Part i,” *Journal of cardiovascular magnetic resonance*, vol. 12, no. 1, p. 71, 2010.
- [221] A. Roche, G. Malandain, X. Pennec, and N. Ayache, “The correlation ratio as a new similarity measure for multimodal image registration,” *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 1115–1124, 1998.

- [222] J. F. Rodríguez-Palomares, M. A. G. Fernández, and J. B. Cosials, “Integrating multimodal imaging in clinical practice: The importance of a multidisciplinary approach,” *Revista Española de Cardiología*, vol. 69, no. 05, pp. 477–479, 2016.
- [223] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, “Svf-net: Learning deformable image registration using shape matching,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 266–274.
- [224] T. Rohlfing, “Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable,” *IEEE transactions on medical imaging*, vol. 31, no. 2, pp. 153–163, 2012.
- [225] O. Ronneberger, P. Fischer, and T. Brox, “U-net: convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*. Springer International Publishing, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_{_}28.
- [226] F. Rousseau, A. D. N. Initiative, *et al.*, “A non-local approach for image super-resolution using intermodality priors,” *Medical image analysis*, vol. 14, no. 4, pp. 594–605, 2010.
- [227] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes, “Nonrigid registration using free-form deformations: Application to breast MR images,” *IEEE Transactions on Medical Imaging*, vol. 18, no. 8, 1999.
- [228] A. Rueda, N. Malpica, and E. Romero, “Single-image super-resolution of brain MR images using overcomplete dictionaries,” *MedIA*, vol. 17, no. 1, pp. 113–132, 2013.
- [229] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *et al.*, “Learning representations by back-propagating errors,” *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.
- [230] C. Rupprecht, E. Huaroc, M. Baust, and N. Navab, “Deep active contours,” *ArXiv preprint arXiv:1607.05074*, 2016.

- [231] M. Saeed, T. A. Van, R. Krug, S. W. Hetts, and M. W. Wilson, "Cardiac mr imaging: Current status and future direction," *Cardiovascular diagnosis and therapy*, vol. 5, no. 4, p. 290, 2015.
- [232] I. S. Salgo, "Three-dimensional echocardiographic technology," *Cardiology clinics*, vol. 25, no. 2, pp. 231–239, 2007.
- [233] F. Sardanelli, M. Quarenghi, G. Di Leo, L. Boccaccini, and A. Schiavi, "Segmentation of cardiac cine MR images of left and right ventricles: Interactive semiautomated methods and manual contouring by two readers with different education and experience," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 785–792, 2008.
- [234] J. Schlemper, J. Caballero, J. V. Hajnal, A. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for mr image reconstruction," in *International Conference on Information Processing in Medical Imaging*, Springer, 2017, pp. 647–658.
- [235] S. Schuler, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *IEEE CVPR*, 2015, pp. 3791–3799.
- [236] S. Schuler, C. Leistner, P. Wohlhart, P. M. Roth, and H. Bischof, "Accurate object detection with joint classification-regression random forests," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2014, pp. 923–930. DOI: 10.1109/CVPR.2014.123.
- [237] *Second Annual Data Science Bowl*, <https://www.kaggle.com/c/second-annual-data-science-bowl>, Accessed: 2017-09-12.
- [238] I. K. Sethi, "Entropy nets: From decision trees to neural networks," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1605–1613, 1990.
- [239] M. Shakeri, H. Lombaert, S. Tripathi, and S. Kadoury, "Deep spectral-based shape features for alzheimer's disease classification," in *Spectral and Shape Analysis in Medical Imaging: First International Workshop, SeSAMI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Revised Selected Papers*. Springer International Publishing, 2016, pp. 15–24. DOI: 10.1007/978-3-319-51237-2{_}2.

- [240] A. Sharma, O. Grau, and M. Fritz, “Vconv-dae: deep volumetric shape learning without object labels,” in *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III*. Springer International Publishing, 2016, pp. 236–250. DOI: 10.1007/978-3-319-49409-8{_}20.
- [241] S. Shernand, “Evaluation of the left ventricular size and function,” in *Comprehensive Atlas of 3D Echocardiography*, Lippincott Williams and Wilkins, 2012.
- [242] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [243] W. Shi, J. Caballero, C. Ledig, X. Zhuang, W. Bai, K. Bhatia, A. de Marvao, T. Dawes, D. ORegan, and D. Rueckert, “Cardiac image super-resolution with global correspondence using multi-atlas patchmatch,” in *MICCAI*, 2013, pp. 9–16.
- [244] W. Shi, X. Zhuang, H. Wang, S. Duckett, D. Oregan, P. Edwards, S. Ourselin, and D. Rueckert, “Automatic segmentation of different pathologies from cardiac cine mri using registration and multiple component em estimation,” in *International Conference on Functional Imaging and Modeling of the Heart*, Springer, 2011, pp. 163–170.
- [245] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, “Real-time human pose recognition in parts from single depth images,” in *CVPR 2011*, IEEE, 2011, pp. 1297–1304. DOI: 10.1109/CVPR.2011.5995316.
- [246] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *ArXiv preprint arXiv:1409.1556*, 2014.
- [247] A. Sotiras, C. Davatzikos, and N. Paragios, “Deformable medical image registration: A survey,” *IEEE transactions on medical imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.
- [248] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *ArXiv preprint arXiv:1412.6806*, 2014.

- [249] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting.,” *Journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [250] S. Standring, *Gray’s Anatomy E-Book: The Anatomical Basis of Clinical Practice*. Elsevier Health Sciences, 2015.
- [251] P. Stolzmann, H. Alkadhi, H. Scheffel, A. Hennemuth, C. Kuehnel, S. Baumueller, S. Kozerke, V. Falk, B. Marincek, and O. F. Donati, “Image fusion of coronary ct angiography and cardiac perfusion mri: A pilot study,” *European radiology*, vol. 20, no. 5, pp. 1174–1179, 2010.
- [252] M. V. Stralen, A. Haak, K. Leung, G. V. Burken, and J. Bosch, “Segmentation of multi-center 3D left ventricular echocardiograms by active appearance models,” *Proceedings of MICCAI CETUS challenge*, pp. 73–80, 2014.
- [253] C. Studholme, C. Drapaca, B. Iordanova, and V. Cardenas, “Deformation-based mapping of volume change from serial brain mri in the presence of local tissue contrast change,” *IEEE transactions on Medical Imaging*, vol. 25, no. 5, pp. 626–639, 2006.
- [254] C. Studholme, D. L. Hill, and D. J. Hawkes, “An overlap invariant entropy measure of 3D medical image alignment,” *Pattern Recognition*, vol. 32, no. 1, pp. 71–86, 1999.
- [255] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, “Multi-view convolutional neural networks for 3D shape recognition,” in *IEEE CVPR*, 2015, pp. 945–953.
- [256] A. Suinesiaputra, B. R. Cowan, J. P. Finn, C. G. Fonseca, A. H. Kadish, D. C. Lee, P. Medrano-Gracia, S. K. Warfield, W. Tao, and A. A. Young, “Left ventricular segmentation challenge from cardiac MRI: A collation study,” in *Statistical Atlases and Computational Models of the Heart. Imaging and Modelling Challenges STACOM 2011*, 2011, pp. 88–97.
- [257] M. Sutton and N. Sharpe, “Left ventricular remodeling after myocardial infarction,” *Circulation*, vol. 101, no. 25, pp. 2981–2988, 2000.

- [258] R. Tanno, A. Ghosh, F. Grussu, E. Kaden, A. Criminisi, and D. C. Alexander, “Bayesian image quality transfer,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp. 265–273.
- [259] R. Tanno, D. E. Worrall, A. Ghosh, E. Kaden, S. N. Sotiropoulos, A. Criminisi, and D. C. Alexander, “Bayesian image quality transfer with cnns: Exploring uncertainty in dmri super-resolution,” *ArXiv preprint arXiv:1705.00664*, 2017.
- [260] G. Tarroni, O. Oktay, W. Bai, A. Schuh, H. Suzuki, J. Passerat-Palmbach, B. Glocker, A. de Marvao, D. ORegan, S. Cook, *et al.*, “Learning-based heart coverage estimation for short-axis cine cardiac mr images,” in *International Conference on Functional Imaging and Modeling of the Heart*, Springer, 2017, pp. 73–82.
- [261] V. Tavakoli and A. A. Amini, “A survey of shaped-based registration and segmentation techniques for cardiac images,” *Computer Vision and Image Understanding*, vol. 117, no. 9, pp. 966–989, 2013.
- [262] M. Toews and W. M. Wells, “Efficient and robust model-to-image alignment using 3d scale-invariant features,” *Medical Image Analysis*, vol. 17, no. 3, pp. 271–282, 2013. DOI: 10.1016/j.media.2012.11.002. arXiv: NIHMS150003.
- [263] A. Torralba, R. Fergus, and Y. Weiss, “Small codes and large image databases for recognition,” in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, IEEE, 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587633.
- [264] P. V. Tran, “A fully convolutional neural network for cardiac segmentation in short-axis MRI,” *CoRR*, vol. abs/1604.0, 2016.
- [265] A. Tsai, A. Yezzi, W. Wells, C. Tempany, D. Tucker, A. Fan, W. E. Grimson, and A. Willsky, “A shape-based approach to the segmentation of medical imagery using level sets,” *IEEE transactions on medical imaging*, vol. 22, no. 2, pp. 137–154, 2003.
- [266] Z. Tu and X. Bai, “Auto-context and its application to high-level vision tasks and 3d brain image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1744–1757, 2010. DOI: 10.1109/TPAMI.2009.186.

- [267] P. Vincent and H. Larochelle, “Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion pierre-antoine manzagol,” *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010. DOI: 10.1111/1467-8535.00290. arXiv: 0-387-31073-8.
- [268] C. Wachinger, M. Toews, G. Langs, W. Wells, and P. Golland, “Keypoint transfer segmentation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9123, 2015, pp. 233–245. DOI: 10.1007/978-3-319-19992-4{_}18.
- [269] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. DOI: 10.1109/TIP.2003.819861. arXiv: 9411012.
- [270] W. Wein, S. Brunke, A. Khamene, M. R. Callstrom, and N. Navab, “Automatic ct-ultrasound registration for diagnostic imaging and image-guided intervention,” *Medical image analysis*, vol. 12, no. 5, pp. 577–585, 2008.
- [271] D. R. Wilson and T. R. Martinez, “The general inefficiency of batch training for gradient descent learning,” *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.
- [272] J. Woo, P. J. Slomka, D. Dey, V. Y. Cheng, B.-W. Hong, A. Ramesh, D. S. Berman, R. P. Karlsberg, C.-C. J. Kuo, and G. Germano, “Geometric feature-based multimodal image registration of contrast-enhanced cardiac ct with gated myocardial perfusion spect,” *Medical physics*, vol. 36, no. 12, pp. 5467–5479, 2009.
- [273] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, “3D ShapeNets: A deep representation for volumetric shapes,” in *CVPR*, 2015, pp. 1912–20.
- [274] S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.
- [275] W. Xue, A. Lum, A. Mercado, M. Landis, J. Warrington, and S. Li, “Full quantification of left ventricle via deep multitask learning network respecting intra-and inter-task relatedness,” *ArXiv preprint arXiv:1706.01912*, 2017.

- [276] Z. Xue, D. Shen, and C. Davatzikos, “Determining correspondence in 3-d mr brain images using attribute vectors as morphological signatures of voxels,” *IEEE transactions on medical imaging*, vol. 23, no. 10, pp. 1276–1291, 2004.
- [277] P. Yan, A. Sinusas, and J. S. Duncan, “Boundary element method-based regularization for recovering of lv deformation,” *Medical image analysis*, vol. 11, no. 6, pp. 540–554, 2007.
- [278] C. Yao, “3D echocardiography image compounding,” PhD thesis, King’s College London, University of London, 2012.
- [279] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, Springer, 2014, pp. 818–833.
- [280] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2010, pp. 2528–2535. DOI: 10.1109/CVPR.2010.5539957. arXiv: 1302.1700.
- [281] Y. Zeng, D. Samaras, W. Chen, and Q. Peng, “Topology cuts: A novel min-cut/max-flow algorithm for topology preserving segmentation in n-d images,” *Computer vision and image understanding*, vol. 112, no. 1, pp. 81–90, 2008.
- [282] 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*, Springer, 2016, pp. 138–145.
- [283] W. Zhang, J. A. Noble, and J. M. Brady, “Spatio-temporal registration of real time 3D ultrasound to cardiovascular MR sequences,” in *MICCAI*, 2007, pp. 343–50.
- [284] W. Zhang, J. A. Noble, and J. M. Brady, “Adaptive non-rigid registration of real time 3d ultrasound to cardiovascular mr images,” in *Biennial International Conference on Information Processing in Medical Imaging*, Springer, 2007, pp. 50–61.

- [285] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Is L2 a good loss function for neural networks for image processing?” *ArXiv preprint arXiv:1511.08861*, 2015.
- [286] X. Zhao, T. K. Kim, and W. Luo, “Unified face analysis by iterative multi-output random forests,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 2014, pp. 1765–1772. DOI: 10.1109/CVPR.2014.228.
- [287] X. Zhen, A. Islam, M. Bhaduri, I. Chan, and S. Li, “Direct and simultaneous four-chamber volume estimation by multi-output regression,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 669–676.
- [288] X. Zhen, Z. 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*, vol. 30, pp. 120–129, 2016.
- [289] Y. Zheng, D. Liu, B. Georgescu, H. Nguyen, and D. Comaniciu, “3d deep learning for efficient and robust landmark detection in volumetric data,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 565–572.
- [290] Y. Zheng, X. Lu, B. Georgescu, A. Littmann, E. Mueller, and D. Comaniciu, “Automatic left ventricle detection in mri images using marginal space learning and component-based voting,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, p. 725 906. DOI: 10.1117/12.811045.
- [291] X. Zhuang, “Multivariate mixture model for cardiac segmentation from multi-sequence mri,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2016, pp. 581–588.
- [292] X. Zhuang, W. Shi, S. Duckett, H. Wang, R. Razavi, D. Hawkes, D. Rueckert, and S. Ourselin, “A framework combining multi-sequence mri for fully automated quantitative analysis of cardiac global and regional functions,” in *Functional Imaging and Modeling of the Heart*, vol. 6666 LNCS, 2011, pp. 367–374.

- [293] M. A. Zuluaga, M. J. Cardoso, M. Modat, and S. Ourselin, “Multi-atlas propagation whole heart segmentation from mri and cta using a local normalised correlation coefficient criterion,” in *International Conference on Functional Imaging and Modeling of the Heart*, Springer, 2013, pp. 174–181.