Politecnico di Torino

EURECOM
Sophia Antipolis

Double Master's degree in *Computer Engineering* (Politecnico di Torino)
Data Science and Engineering (EURECOM - Télécom Paris)

Master's Thesis

Cardiac Image Segmentation: towards better reliability and generalization

Supervisors:
GARZA Paolo, Politecnico di Torino
ZULUAGA Maria A., EURECOM

Candidate: **Francesco Galati**

March-April 2021

Ai miei nonni
Armando, Armando,
Maria e Teresa.

Al professore Roberto.

Summary

Cardiac image segmentation is the problem of learning the anatomical semantics of each voxel in a three-dimensional heart image. In clinical practice, radiologists are delegated to draw contours manually, encompassing the structures of interest. The process is lengthy, monotonous, and prone to subjective errors. Starting from the 1970s, researchers have thoroughly investigated the possibility of automating this task. Automated CMR segmentation can help clinicians interpreting the medical conditions, speeding up diagnoses, increasing monitoring reliability, facilitating surgical planning, and enabling vast population studies. Overall, it would make a strong contribution to the battle against cardiovascular diseases (CVDs), estimated to cost 31% of all global deaths. During the last decade, this automation attempt has been lead by deep learning.

Between 2013 and 2015, deep learning techniques became popular, and more and more papers on the topic went public. When the MICCAI conference of 2017 hosted the ACDC Challenge, nine participants out of ten implemented a deep convolutional architecture to fulfill the segmentation task. This brief time window represents a drastic change in the field. Results reveal that deep learning methods can successfully classify patient data and get highly accurate segmentation results. However, these approaches require fully annotated datasets, which must capture the anatomical variability of heart images. Collecting so much data requires extensive human effort. In addition, neural networks do not naturally provide probabilistic guarantees on their predictions. The inclusion of an external monitoring mechanism is crucial to ensure the reliability of subsequent diagnoses.

This thesis attempts to solve both the problems of generalization and automatic quality assessment. The proposed solutions revolve around the development of a convolutional autoencoder, which provides a surrogate quality measure for individual segmentation masks and their generating model. In particular, we propose two different types of measures, a global score, and a pixel-wise map, and we demonstrate their use by reproducing the results of the ACDC Challenge in the absence of ground truth. Next, we integrate our autoencoder into a semi-supervised framework, capable of learning from both labeled and unlabeled data to fulfill the segmentation task.

Contents

1 Introduction

2 Background

2.1 Anatomy of the heart . . .
2.2 Cardiovascular diseases . . .
2.3 Cardiac imaging . . .
2.4 CMR Segmentation . . .

3 Efficient model monitoring for quality control in cardiac image segmentation

3.1 Motivation . . .
3.2 Method . . .
3.2.1 Background . . .
3.2.2 Model Monitoring Framework . . .
3.2.3 Network Architecture . . .
3.3 Experimental Setup . . .
3.3.1 Data . . .
3.3.2 Setup . . .
3.3.3 Implementation . . .
3.4 Results . . .
3.5 Conclusions . . .

4 Semi-supervised segmentation for improved cross-domain generalisation

4.1 Motivation . . .
4.2 Method . . .
4.2.1 The segmenter $\mathcal{M}$ . . .
4.2.2 The image reconstructor $\mathcal{R}$ . . .

4.2.3 Two-phase training . . . . . .
4.2.4 Implementation. . . . . .
4.3 Experiments and Results . . . . . .
4.3.1 Experimental Setup . . . . . .
4.3.2 Results . . . . . .

5 Conclusions

Bibliography

Chapter 1

Introduction

This thesis investigates the state-of-the-art in cardiac image segmentation, which is the problem of learning the anatomical semantics of each voxel in a three-dimensional heart image. Our first reference is the ACDC Challenge [1], hosted in 2017 at the MICCAI conference. On that occasion, it was released a dataset of 100 healthy and non-healthy patients from the University Hospital of Dijon (France). In a total of ten participants, nine implemented deep learning techniques, which achieved human performances in the segmentation of the left ventricle, the myocardium, and the right ventricle. The proposed models, however, suffer from significant weaknesses, which we examine and attempt to solve.

Chapter 3 focuses on automatic quality assessment, which consists of identifying unusable segmentations, ensuring the reliability of subsequent diagnoses. Indeed, failures can occur in both the acquisition process and the segmentation process of CMR images. In clinical practice and population studies, it is of utmost importance to develop an automatic supervision system placed alongside the segmentation model, guaranteeing its safe use even in the absence of ground truth. Previous approaches require manual annotations to estimate segmentation performances, which can be difficult to obtain, or require spatial alignment between ground truth images and segmentation, achieved by the use of image registration. With this in mind, we propose a novel learning framework, which addresses the limitations of past techniques thanks to its formulation under an anomaly detection paradigm.

Chapter 4 focuses on the problem of generalization, which consists of accurately predicting outcome values for previously unseen data. The models trained on the ACDC dataset do not perform so well when tested on images taken with different scanners or protocols or depicting peculiar heart deformations. A first viable option includes enough anatomical and perspective variability in a unique dataset to represent a large slice of the population, e.g., the UK Biobank dataset [2]. However, the manual delineation of important structures within the cardiac image typically takes several minutes, even for a trained expert. This represents a bottleneck towards large scale data collection. As an alternative, previous works attempt to solve the problem by decreasing the model complexity or increasing the amount of training data. On this second path, we propose a novel semi-supervised framework to include unlabeled data for training.

Chapter 2

Background

2.1 Anatomy of the heart

The heart is one of the most complex and essential organs in the human body, as well as in most animals. With its 75 beats per minute (and an average of three billion heartbeats over a lifetime), it provides pressure to the circulatory system. In this way, the heart allows the approximate 5 liters of blood in an adult body to flow inside a 100,000-kilometer-long network. The blood brings oxygen and important nutrients to the body's tissues and organs through the arteries, and it carries metabolic waste on its way back through veins. In humans, the heart is approximately the size of its owner's closed fist and is located underneath the sternum and ribcage, between the lungs, in the middle compartment of the chest.

The heart is divided into four chambers functioning as a double-sided pump, with an upper atrium and a lower ventricle on each side of the heart [Figure 2.1]. The right atrium and ventricle are referred to together as the right heart and their left counterparts as the left heart. The four chambers are divided by a wall of muscle called the septum. More in detail, the atrioventricular separates the atria from the ventricles, the interatrial septum separates the atria and the interventricular septum separates the ventricles. Finally, the heart has four valves, one between each atrium and ventricle, and one at the exit of each ventricle: the tricuspid valve, between the right atrium and the right ventricle, the mitral valve, between the left atrium and left ventricle, the pulmonary valve, at the exit of the right ventricle, and the aortic valve, at the exit of the left ventricle.

The heart's atria receive blood from the veins, while the heart's ventricles pump blood into the arteries. In particular, when working properly, oxygen-depleted blood coming from the body's tissues and organs, except for the lungs, enters the right atrium. The blood then passes through the tricuspid valve into the right ventricle, which pumps it to the pulmonary artery. Through this artery, the blood reaches the lungs, where, during air exchange, it receives oxygen in exchange for carbon dioxide. The oxygenated blood returns to the heart through the left atrium, it flows through the mitral valve into the left ventricle and is finally pumped back to the body through the aorta. Being responsible for reaching every end of the

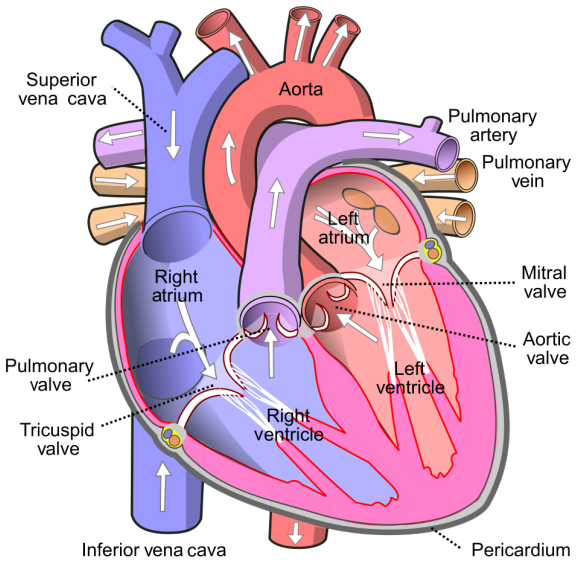

Figure 2.1. Anatomy of the heart.

body other than the lungs, the left ventricle has the thickest muscle mass of all the chambers.

Atria and ventricles contract to make the heartbeat and pump the blood. Each beat counts two phases: diastole and systole. Systole occurs when the heart contracts to release blood, diastole follows as the heart relaxes. The contractions of the cardiac muscle are triggered by involuntary electrical pulses coming from the brain, which keep the blood flowing in proper rhythm. This muscle is called the myocardium, and together with the inner endocardium and the outer epicardium, forms the heart wall.

2.2 Cardiovascular diseases

Cardiovascular diseases (CVDs) are estimated to cause 31% of all global deaths, the highest rate among all the pathologies. CVDs include vascular diseases, which involve the blood vessels, and heart diseases, which indicate in more specific terms pathologies affecting the heart. Vascular diseases compromise the circulatory system, and they can cause different signs and symptoms all over the body. Since this thesis is specifically focused on cardiac imaging, speculating on that large part of vascular diseases that do not leave any trace on the heart is beyond the scope of this work.

Among vascular diseases, Coronary Artery Disease (CAD) is worth to mention, since it obstructs the arteries that move oxygen-rich blood through the heart and the lungs, and can lead to terrible consequences (e.g., stroke, heart failure), as well as to cardiac deformities (e.g., enlargement of the left ventricle or right ventricle). If deformities are long-term repercussions, heart failure can be promptly detected

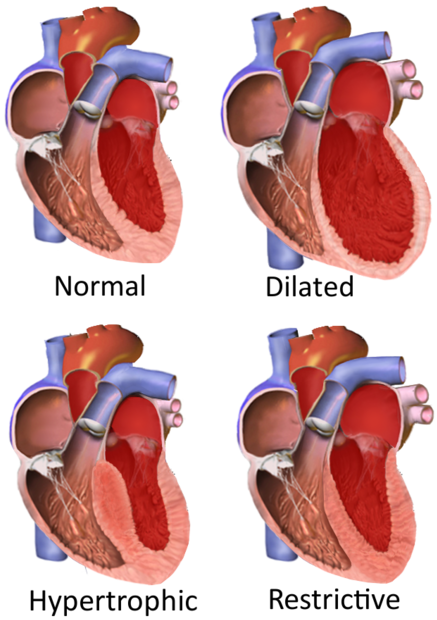

Figure 2.2. Structural categories of cardiomyopathy.

in cardiac images by calculating the Ejection Fraction (EF). This measure reflects how much blood leaves a heart ventricle every time it pumps. The EF is expressed as the percentage of blood pushed out from the left ventricle over the total amount of blood in it. Values between 50% and 70% are considered normal, while a value below 40% is an indicator of systolic heart failure. Heart patients can also show another type of heart failure, called diastolic heart failure, which occurs when the left ventricle contracts normally during systole, but the ventricle is stiff and does not relax normally during diastole, which impairs filling.

Heart diseases can lead to deformations. For example, cardiomyopathy is a group of pathologies that cause the heart muscle to grow larger and turn rigid, thick, or weak. Types of cardiomyopathy include hypertrophic cardiomyopathy, dilated cardiomyopathy, restrictive cardiomyopathy, arrhythmogenic right ventricular dysplasia, and Takotsubo cardiomyopathy [Figure 2.2]. Dilated cardiomyopathy originates from a general enlargement of the heart, which cannot pump blood effectively. Patients with this disease have an ejection fraction below 40%, and a large left ventricular volume. Hypertrophic cardiomyopathy, instead, involves the heart's walls, which become thicker. This condition does not compromise the cardiac function, which measures an ejection fraction greater than 55%. Finally, restrictive cardiomyopathy causes the stiffening of the heart's wall, without any thickening. Thus the heart is restricted from stretching and filling with blood properly. The last ones especially noteworthy are congenital heart defects, which indicate irregularities that are present at birth. Some of these defects are never diagnosed, others may be found when they cause symptoms.

2.3 Cardiac imaging

Having mentioned congenital heart defects, genetic factors may have an impact, but these diseases are often caused by poor lifestyle habits, such as poor diet, lack of regular exercise, tobacco smoking, alcohol or drug abuse, and high stress. These issues are prevalent in modern Western culture, but heart disease has always plagued the human race. Studies have shown that even ancient Egyptian mummies had identifiable cardiovascular diseases, specifically, atherosclerosis in different arteries of the body [3]. However, until the discovery of X-rays in 1895, there were no modalities allowing clinicians to look at the heart, and diagnoses were all given by physical examination and a doctor's best guess. From 1896, radiographic findings for heart disease began to be characterized and recorded. With the support of radiography, clinicians could assess the size of the heart chambers by looking at the different silhouettes of the cardiac shadow.

In the last century, cardiac imaging has come a long way, witnessing extraordinary advances in the capacity to display interior and borders of a living human's heart [4]. Nowadays, many cardiovascular imaging modalities allow evaluating the heart condition, such as echocardiography, Cardiovascular Computerized Tomography (CCT), Cardiovascular Magnetic Resonance Imaging (CMR, also known as cardiac MRI), invasive coronary angiography, cardiac Positron Emission Tomography (PET), and Nuclear Cardiology (NC).

This thesis focuses on CMR, an imaging technology for the non-invasive investigation of cardiovascular diseases, which started to develop in the 1970s. General utilization of MR makes strong use of magnetic fields and radio waves by observing the polarity due to hydrogen nuclei spin to generate images of the organs in the body. Subsequently, the alignment of this magnetization is changed by emitting radio frequency pulses which produce a rotating magnetic field detectable by an external RF coil. Owing to the beating and breathing motions of the heart, conventional MRI sequences were adapted for cardiac imaging by introducing ECG gating in 1983, which allows for stop motion-imaging by acquiring data only during a specified portion of the cardiac cycle, typically during diastole when the heart is not moving. Increasingly sophisticated techniques were developed, including cardiac cine-MRI, considered today the standard technique for achieving high resolution and evaluating global function measurements through segmentation, and tagged MRI, which uses spin tagging prepulse to produce detectable markers over time, permitting regional analysis and temporal motion registration. The development of CMR is an active field of research and continues to see a rapid expansion of new and emerging techniques. Since MRI uses non-ionizing radiation, it is considered a non-invasive technique. Advantages of CMRs include the possibilities of well visualizing the myocardium, acquiring images with different orientations, and of evaluating perfusion, function, scars, and epicardial coronaries. MR images acquired with an orientation perpendicular to the long axis of the heart is called the Short Axis plane (SAX) [Figure 2.3]. CMRs generally cover about 10–15 z-plane slices and 15–30 temporal frames, depending on the size of the heart. However, this technology is more expensive than other methods, and it is not always available in

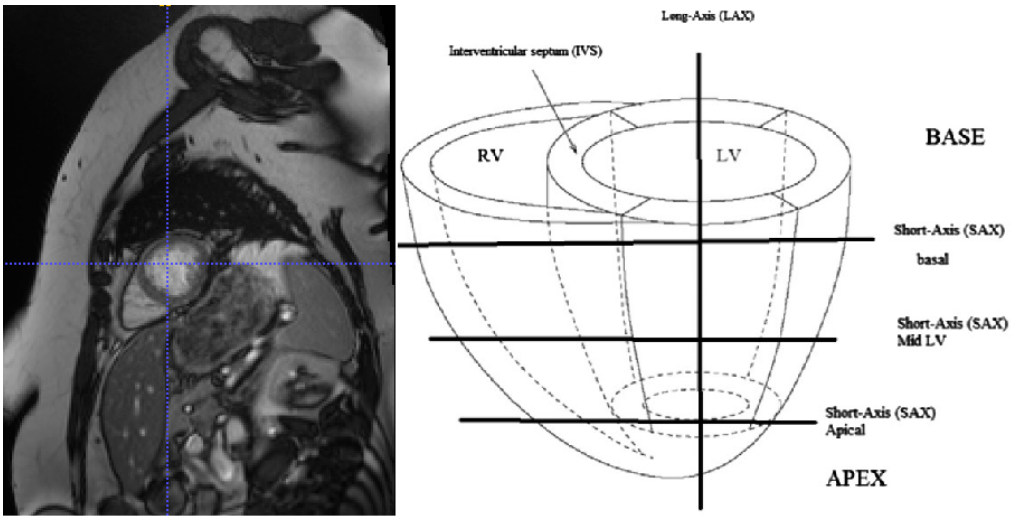

Figure 2.3. Cardiac Cine CMR SAX on the left, illustration of Short-Axis (SAX) and Long-Axis (LAX) cardiac images on the right.

cardiac care centers. Moreover, pacemakers and metal implants, often present in heart patients, are contraindications to the use of MRI.

2.4 CMR Segmentation

When dealing with CMRs, doctors are generally interested in measuring clinical parameters such as ventricular volumes, myocardial mass, and ejection fraction. The primal step consists of delineating important organs and structures within the cardiac image: a radiologist is usually delegated to manually draw contours encompassing the structures of interest, an analysis which typically takes a trained expert around 20 minutes per subject. This is lengthy, monotonous, and prone to subjective errors. For these reasons, researchers have thoroughly investigated the possibility of automated CMR segmentation and analysis to immediately trace the borders of the main subparts in a given heart image. Historically, an initial focus was oriented towards the left ventricle. This choice is motivated by increased variability and concavity in the shape of the RV, whose segmentation becomes more challenging than the LV. Apical slices seem also to be more difficult to segment due to less and unpredictable information at the end of LV and RV.

From the 1970s to the 1990s, low-level pixel processing and mathematical modeling were used together to construct rule-based systems for medical image analysis. Among these techniques, thresholding was one of the simplest and most popular. It can be performed exploiting both global (e.g., the gray level histogram of the entire image) or local (e.g., co-occurrence matrix) information. In the straightforward case of an image composed of two distinct regions with different gray level ranges, its histogram will show two peaks, separated by a valley. The bottom of the valley is then taken as a threshold for object background separation. However, in realistic cases, threshold selection is not such a trivial and effective job. More complex variants of this method select multiple thresholds to achieve heterogeneous

image segmentation (i.e, adaptive thresholding) [5, 6, 7] or multi-class segmentation (i.e, multithresholding) [8, 9]. Sahoo [10] and Lam [11] provided a wide review on thresholding techniques. The latter, in a wider perspective, summarized other popular techniques of those times. For example, relaxation [12, 13] is an iterative approach which allows to classify each pixel in parallel. At each iteration, decisions made at neighboring points are combined to decide on the next iteration. Another popular example to be cited is edge detection. In a grayscale image, an edge is a set of points of abrupt changes in intensity values. Being low-level features, edges can be detected only basing on local information, without any high-level comprehension of the image. This intuition points towards parallel solutions, which determine whether a point is an edge or not basing few neighboring points. In general, a large variety of parallelization levels characterize the methods available in the literature [14].

The pixel-driven techniques described above, however, turned out to be incapable of detecting object boundaries in cardiac texture, due to their wide intensity range and their lack of overlaps on the edges. This problem was tackled at the end of the 1990s by introducing statistical and physical assumptions derived from the cardiac structure. Supervised techniques, making use of a set of training data as examples or templates, became popular to build statistical models. Instead of relying exclusively on the intensity values to draw contours in cardiac images, statistical techniques define additional shape, motion, or texture (intensity variations) priors to constrain the delineation process. The reasons behind this intuition lie in the fact that heart contours do not change dramatically in the spatial and temporal direction, as well as among different patients. In general, all human hearts can be modeled as ellipsoidal objects moving in- and out-wards. From this perspective, several authors tried to build cardiac priors and integrate them into different segmentation techniques as an additional constraint to overcome the failures of low-level pixel methods. Statistical methods can be classified according to the way they define and make use of priors, which allow the manipulation of previously segmented contours to locate new ones. Active Shape Models (ASM) [15, 16, 17] build a shape prior starting from manually segmented data. A shape model reflects the typical structure of a set of anatomical objects, therefore being invariant to transformations applied on them. The segmentation process starts by aligning all the training data to a defined coordinate using rigid registration techniques. In the case of a heterogeneous number of frames, an interpolation in time is performed to make it equal. Given then the set of coordinates, the shape model can be defined as a statistical map of these points in the form of mean $\bar{\mathbf{x}}$ and covariance $\mathbf{S}$:

$$\bar{\mathbf{x}} = \frac{1}{N}\sum \mathbf{x_i} \tag{2.1}$$

$$\mathbf{S} = \frac{1}{N-1}\sum (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T \tag{2.2}$$

The principal eigenvectors are then extracted from the covariance matrix. Each training point can be approximated as:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{2.3}$$

where $\mathbf{P}$ is the matrix collecting the selected eigenvectors, and $\mathbf{b}$ is a vector of weigths. By varying these parameters within suitable limits, new examples of the same shape can be generated. Finally, segmentation of a new object $\mathbf{x}'$ is performed by overlapping the statistical model on it and estimating the transformation T that leads to maximal correspondence:

$$\mathbf{x}' = T(\bar{\mathbf{x}} + \mathbf{P}\mathbf{b}) \tag{2.4}$$

For all the steps described above several variations can be found in the literature. The interested readers are referred to [18] for a further study on the alternatives. For example, multi-atlas techniques [19, 20, 21, 22] dispose of multiple options instead of defining a single prior representing the whole population. Active Appearance Models (AAM) [23, 24, 25] follow the same procedure as ASM, but including the texture appearance of the object as well. Finally, motion models [26, 27] take into account the cardiac cycle to build a motion prior as a temporal constraint.

During the same years, Machine Learning (ML) started to be deployed for diagnostic purposes. The typical ML approach consisted of extracting features from cardiac images with both manual or automatic techniques (such as the segmentation techniques discussed above) to build a statistical classifier capable of predicting the presence or absence of specific diseases. This pattern brought to the development of systems progressively shifting to complete automation. Indeed, handcrafted features were initially extracted from the medical images and fed into automatic classifiers. With the advent of deep learning, computers were programmed to learn by themselves the optimal features for classification. Especially in the case of images, Convolutional Neural Networks (CNNs) have succeeded in learning increasingly higher-level features to turn input images into spatially comprehensive outputs.

CNNs have demonstrated massive efficiency in automating time-consuming visual tasks, often surpassing human performances in several research fields, such as medical imaging, video surveillance, and autonomous driving. In this thesis, driven by incentive past results, the clinical practice described above becomes the subject of further investigations regarding automatic cardiac image segmentation. Thanks to the exponential growth of computer power along with the availability of public databases, several works in recent years have already explored the possibility of developing a computer model capable of accomplishing CMR segmentation with a nearly human performance at a relatively low cost.

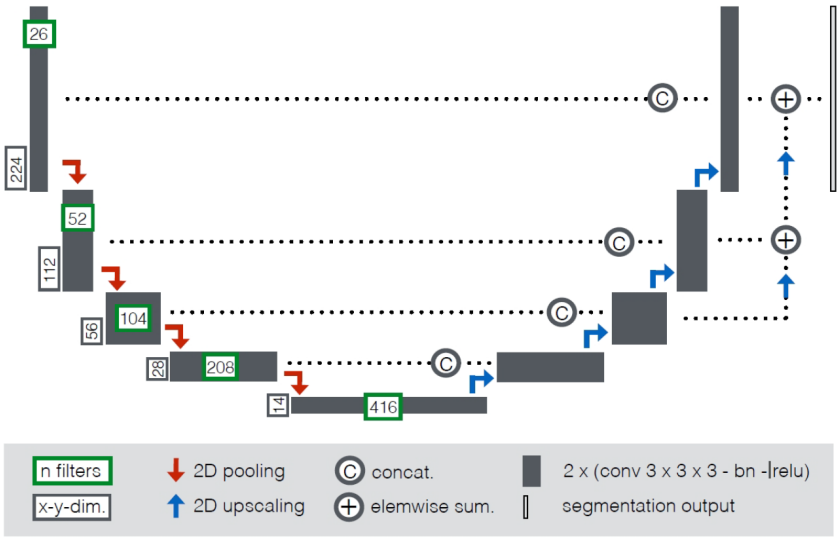

Figure 2.4. Architecture developed by Isensee [28]. The image portrays the 3D network. The 2D network is equivalent, but uses 2D convolutions, patch size 352×352 and 48 initial features.

In 2017, the MICCAI conference hosted the ACDC Challenge [1], whose goal was to achieve human performances in the segmentation of the left ventricle, the myocardium, and the right ventricle. On that occasion, it was provided training set with 100 subjects, each paired with the corresponding annotation given by one clinical expert. Manual references allow performing supervised image segmentation, which aims to learn the semantics of each pixel by minimizing a global dissimilarity measure between the ground truth and the segmentation result. After being trained, such a model will be capable of labeling each pixel with a different color referring to the part of the heart it belongs to. Nine research groups developed deep learning methods to fulfill the segmentation task, and the top finishers of the challenge showed that CNNs could successfully get highly accurate segmentation results.

Isensee [28] developed an ensemble model of 2D and 3D U-Net [29, 30] inspired architectures, integrating segmentation and disease classification into a fully automatic processing pipeline. Due to different memory requirements, the two architectures present some differences. First of all, they adopt two different convolution operations, the 2D convolution and the 3D convolution. In the 2D model, the number of initial feature maps is 48, and the input patch size is 352×352. For the 3D model, these values drop to 26 and 224×224 respectively. Figure 2.4 shows the layers and the connections which characterize the networks. Each feature extraction block consists of two padded convolutions, followed by batch normalization and a leaky ReLU nonlinear activation. The number of feature maps is then doubled (halved) with each of the 4 pooling (upscaling) operations. Such operations are carried out only in the short axis plane for the 3D network. Finally, residual connections are inserted along with the upsampling layers. The networks are trained using a multiclass dice loss. Before each of the last two upscaling operations, deep supervision is implemented by generating low-resolution segmentation outputs via

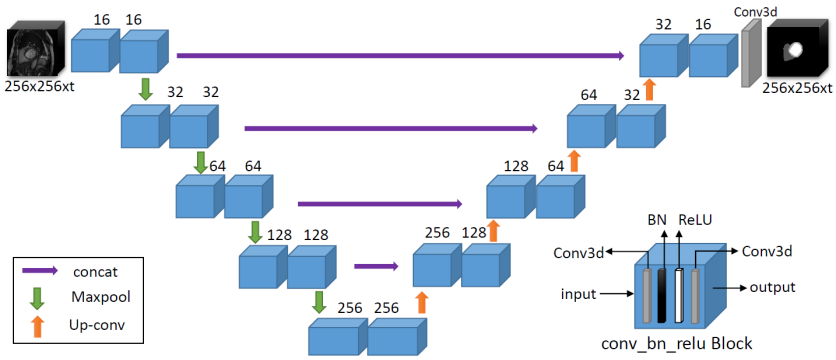

Figure 2.5. Architecture developed by Patravali [33]. The image portrays the 3D network.

1x1x1 convolutions, which are then upscaled and aggregated for the final segmentation. To increase the generalization, a broad range of data augmentation techniques was adopted, such as mirroring, random rotations, gamma-correction, and elastic deformation. Due to the low z-resolution, all data augmentation was performed only in the x-y plane. All slices within the training batch were perturbed with a probability of 10% and a random offset drawn from N(0, 20) to account for the presence of slice misalignments.

Baumgartner [31] investigated the impact of using 2D and 3D convolutional layers, as well as using different losses. In particular, the paper cover 4 convolutional neural network architectures: FCN-8 [32], 2D U-Net, modified 2D U-Net, and modified 3D U-Net. The modified versions set the number of feature maps in the transpose convolutions of the upsampling path equal to the number of classes, intuitively associating each class with at least one channel. Both versions of the 2D U-Net-based models outperform FCN-8 and the 3D U-Net, while of the two the modified version leads to slightly better results. The authors highlight the possible reasons, as the reduction of the total amount of parameters when using 2D U-Net networks, the reduction of the total amount of training samples when using 3D data, the low resolution of the through-plane, and the substantial downsampling of 3D data due to GPU memory restrictions. All the networks were trained with 3 different loss functions: the standard pixel-wise cross-entropy, a weighted pixel-wise cross-entropy to account for the class imbalance between the back-ground and the foreground classes, and finally the dice loss. The weighted loss function led to marginally better results than the standard cross-entropy, both surpassing the dice loss.

Similar to [31], Patravali [33] tested 2D and 3D U-Net based networks with different Dice and cross-entropy losses. The architecture is shown in Figure 2.5. Each block (in blue) consists of two convolutions, intercalated with batch normalization and a ReLU activation. The 2D segmentation model is trained slice-by-slice, whereas we compute volumetric segmentation for the 3D model. In order to remove inconsistencies in the dataset and ensure that the model receives uniform inputs, preprocessing steps are performed, such as CLAHE to remove noise and enhance contrast [34], normalization, clipping, resampling to a common voxel spacing of

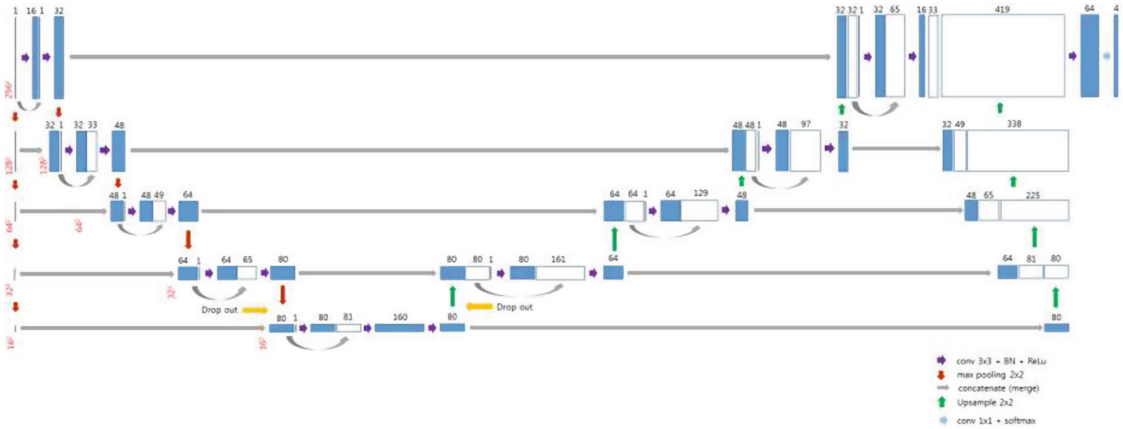

Figure 2.6. Architecture developed by Jang [36].

1.5x1.5x10, resize and crop to a fixed size of 256x256. Finally, on a random basis, the data is rotated between -15 to +15 degrees and scaled between 0.9 - 1.1 range to ensure slight robustness and variability in training the network. All the networks are trained with 3 possible loss functions, the Cross-Entropy Loss (CE Loss), the Dice Loss, and the Combined Cross Entropy-Dice Loss (Dice-CE Loss). From the experiments, the proposed dice loss function outperforms CE Loss and CE-Dice Loss functions across all metrics in both 3D and 2D models.

Yang [35] implemented a 3D U-Net, which substitutes the usual concatenations with residual connections to smooth gradient flow. The training process conducts transfer learning starting from a pretrained C3D model, and it exploits a deep supervision mechanism, attaching several auxiliary loss functions to expose early layers to better supervision. Limited by volume dimension, only 2 pooling layers are inserted. Each convolutional layer is followed by a batch normalization layer and a rectified linear unit (ReLU). The network is trained with a multi-class Dice loss. Before feeding the data into the network, the intensity of original MR volumes is calibrated with the CLAHE algorithm and normalized as zero mean and unit variance. Random rotation with a probability of 0.30 is used to augment the training dataset. The third dimension of the volume is resized to 32 during training and testing to facilitate consecutive 3D convolution and pooling.

Jang [36] implemented an M-Net [37], which adds to the standard U-Net architecture some concatenations between feature maps of the adjacent layers. The network is trained end-to-end from scratch. The authors observe that, since the training dataset has a relatively large slice thickness (from 5 mm to 10 mm), 3D information degrades the performance and impedes generalization of the model, making 2D convolutions preferable. Therefore, the proposed FCN architecture [Figure 2.6] has the same layers with M-net excluding the 3D convolution filter. Due to the large variety of in-plane dimensions, from 154×224 to 427×512, the authors chose to re-scale the maximum size of each image to 256, padding the residual regions accordingly to get constant width and height of 256×256. The wide range of voxel intensity resulting from the use of different scanners or acquisition protocols is normalized by subtracting to the voxel intensity of each image the mean and

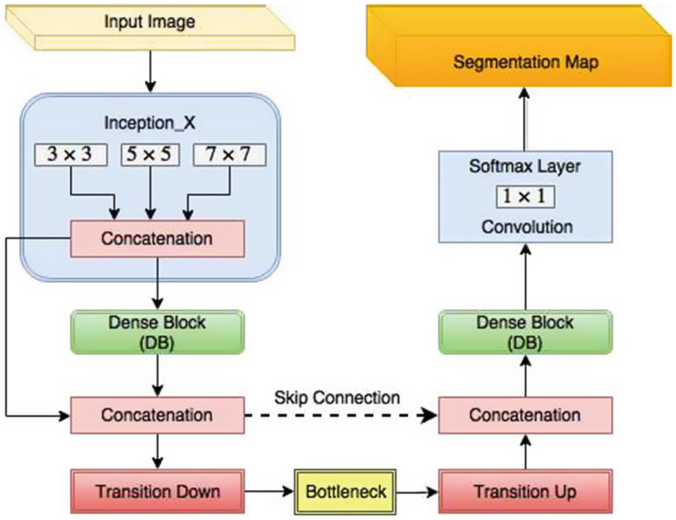

Figure 2.7. Architecture developed by Khened [38].

dividing it by its standard deviation. Data augmentation is performed by rotating each image from -60 to +60 degrees at uniform intervals of 15 degrees. A convex hull is also applied to remove concavities only for LV. Finally, in order to balance the contributions of each class to the training loss, a weighted cross-entropy is used.

Khened [38] developed a dense U-Net. Dense CNNs facilitate multi-path flow for gradients between layers, and significantly reduce the number of parameters, which is ideal to avoid overfitting in scenarios with a limited amount of data. Since the cardiac MR images of each patient portray various surrounding structures (e.g., the lungs and diaphragm), they are subject to a huge class imbalance concerning the background. To alleviate this problem, their method extracts the Regions of Interest (ROI) by applying a Fourier and a Canny edge detector, followed by a circular Hough transform to compute an approximate radius and center of the LV. Starting from the center, patches of size 128×128 are extracted from each scan. The proposed technique enables the network to precisely learn the fine-grained structures of the heart while reducing the computation time required for learning the parameters of the network and also during inference. In order to augment the training data, rotations, translations, rescaling, and flipping operations are employed. Before feeding the data into the model, the voxel intensities of each CMR image are normalized to the range of 0-1. The architecture of the network is shown in Figure 2.7. The down-sampling and up-sampling components of the network adopts the fully convolutional DenseNets architecture for semantic segmentation [39]. Dense blocks concatenate new feature maps created at a given resolution. The use of dense blocks instead of basic convolution blocks makes the system lighter, passing from a total number of trainable parameters equal to 30M in the U-Net to 4M. Each layer in the dense block is sequentially composed of Batch Normalization, Exponential Linear Unit (ELU, observed to make the system converge faster than the ReLU), a 3×3 convolution, and a dropout of 0.2. The first layer corresponds to an inception layer [40], which concatenates 3×3, 5×5 and 7×7 convolutions. The Transition-Down block (TD) implements a 1×1 convolution and a 2×2 max-pooling layers. A skip

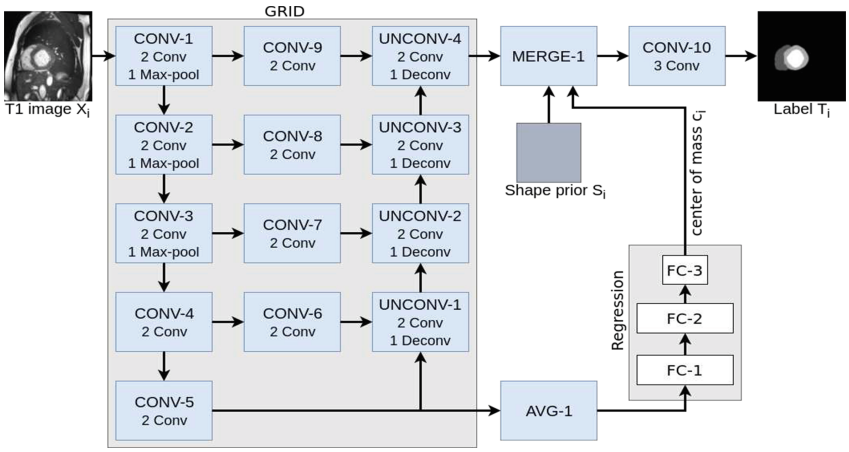

Figure 2.8. Architecture developed by Zotti [43].

connection concatenates the output feature maps of the TU block and the output of the DB block. Finally, the last layer implements a 1×1 convolution layer followed by a soft-max operation to generate the final label map of the segmentation. The parameters of the network were optimized by training with a weighted dual cost function, the sum of the Dice loss and the Cross-Entropy loss.

Rohé [41] implemented a multi-atlas segmentation framework. The characteristic of such a framework is to divide the workflow into two phases: registration phase and fusion phase. The first consists of overlaying the target image with all the images in the training set, in order to create a geometrical alignment, useful for further analysis. The registration step consists of rigid and non-rigid steps. It is important to notice that the position and orientation of the heart in the target image must be known prior to the registration step, in order to allow an appropriate alignment. This task is performed with a CNN trained to detect two landmarks. Given this information, the rigid registration step becomes a trivial alignment of the images, while the non-rigid step relies on SVF-Net [42]. The fuse phase consists of a soft fusion method that merges the registered label fields, using pixel-wise confidence measures.

Zotti [43] implemented an architecture, called GridNet, which is specifically designed to segment the CMR images, embedding a shape prior and a loss function tailored to the cardiac anatomy. The model integrates a cardiac center-of-mass regression module and a segmentation module. The regression module, together with a precalculated shape prior S, allows for an automatic shape prior registration by translating S on the estimated center of mass. The shape prior S is a 3D volume that encodes the probability of each voxel being part of a certain class. Such probability is estimated by computing the pixel-wise empirical proportion of each class based on the ground truth label fields provided with the training dataset. The architecture is made of a grid-like CNN network with 3 columns and 5 rows [Figure 2.8]. The authors assert that a common issue with MRI cardiac images is the fact that along the 2D short-axis, the location of the heart sometimes gets shifted from one slice to another due to different breath-holds during successive

acquisitions. For this reason, the network is fed exclusively with 2D slices taken from the ED or ES phase independently, on which 2D convolutions of filter size 3×3 are performed. The first column of convolutions (from CONV-1 to CONV-5) extracts high-level features, used to predict the cardiac center of mass. The second column (from CONV-6 to CONV-9) contains 4 convolution layers used to compute features at various resolutions. The last column (from UNCONV-1 to UNCONV-4) aggregates features from the lowest to the highest resolution, merging both global and local features used to segment the image. Overall, the architecture can be seen as an extension of the U-Net, except for the middle CONV-6 to CONV-9 layers along with the skip connections and for the use of CONV-5 for center-of-mass estimation. Finally, the MERGE-1 layer is fed with 7 feature maps: 4 coming from UNCONV-4 and 3 from the realignment of S, based on the estimated center of mass. The architecture makes use of batch normalization, the ReLU activation function, and dropout. The number of total parameters decreases from 32 million in the original U-Net to 8 million. The loss used to train the GridNet incorporates four terms: the cross-entropies of the predicted labels and the predicted contours, the Euclidean distance between the predicted and the true centers of mass, and the prior loss. The authors also precise they make no use of any manual preprocessing and image cropping so that their model learns both high-level features (useful to distinguish the heart from other organs with a similar shape) and low-level features (useful to get accurate segmentation results).

Wolterink [44] developed a CNN without an encoder-decoder architecture. At first, to normalize the differences in voxel size, all 2D images are resampled to 1.4×1.4 mm^2 spacing. To correct for image intensity differences between images, each MR volume is rescaled between 0 and 1 according to the 5th and 95th percentile of intensities in the image. The network is designed to contain a sequence of convolutional layers with increasing levels of kernel dilatation to predict each pixel's label by relying on a sufficiently large receptive field. The model counts two separated input channels for the ED and the ES phases, and eight output channels, one per class and phase. The trainable parameters of the CNN are optimized by minimizing a soft Dice loss function. Potential overfitting of the network was mitigated by the inclusion of Batch Normalization layers and of L2-regularization.

Tziritas and Grinias [45] implemented a Chan-Vese level-set method followed by an MRF graph cut segmentation method and spline fitting to smooth out the resulting boundaries. Since this method does not implement a deep convolutional architecture, the details are omitted since such techniques are out of the scope of this thesis work.

Few months after the ACDC Challenge, similar highly accurate results were achieved by Bai [46] on a far larger dataset, acquired from the UK Biobank [2]. The dataset collects CMR images of 5008 patients, each pixel-wise annotated by a team of 8 experts. The proposed model [Figure 2.9] is adapted from the VGG-16 and it consists of 16 convolutional layers with 3×3 kernel, followed by batch normalization and a ReLU activation. After every two or three convolutions, the feature map is downsampled by a factor of 2 so as to learn features at a more global scale. Feature

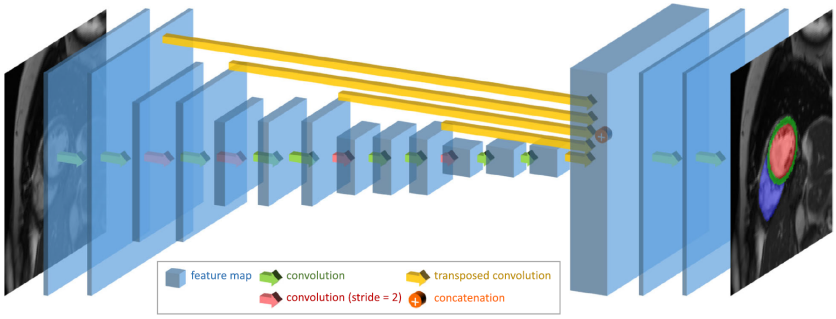

Figure 2.9. Architecture developed by Bai [46].

maps learned at different scales are then upsampled to the original resolution using transposed convolutions to be then concatenated. Finally, three convolutional layers of kernel size 1×1, followed by a softmax function, are used to predict a probabilistic label map. The mean cross-entropy between the probabilistic label map and the manually annotated label map is used as the loss function. This architecture differs from U-Net for the upsampling half of the network, which iteratively doubles the feature map at each scale. Before feeding the images to the network, they are all cropped to the same size of 192×192 and intensity normalized between 0 and 1. Data augmentation is performed on-the-fly by applying random translation, rotation, scaling, and intensity variation to each mini-batch of images.

The authors tested their model on both the ACDC dataset and the UK Biobank dataset. Their expectations were unfulfilled: once trained on one dataset, the model showed low performance on the other. This pointed out a clear generalization problem for cardiac image segmentation.

Poor generalization is not the only contemporary problem in CMR segmentation. The current bottleneck towards the large scale use of learning-based pipelines comes from the monitoring and maintenance of the deployed ML systems. In clinical practice, a quality control step comes right after the acquisition through visual inspection. In the case of effective automation of the only segmentation process, the result is subject to a careful quality check by a human operator, and thus not fitting into high-throughput acquisition protocols. However, the identification of unusable segmentations is crucial to ensure the reliability of subsequent diagnoses, and consequently, it must be automized as well and not left behind.

The problem of automatic quality assessment and the problem of generalization are the protagonists of the next two chapters.

Chapter 3

Efficient model monitoring for quality control in cardiac image segmentation

In this chapter, we present a novel learning framework to monitor the performance of cardiac image segmentation models in the absence of ground truth. Formulated as an anomaly detection problem, the monitoring framework allows to derive surrogate quality measures for a segmentation and allows to flag suspicious results. The intuition behind this work lies on the possibility of estimating a model of variability of cardiac segmentation masks from a reference training dataset provided with a reliable ground truth. This model relies on a a convolutional autoencoder, which can be subsequently used to identify anomalies in segmented unseen images. We propose two different types of quality measures, a global score and a pixel-wise map. We demonstrate their use by reproducing the final rankings of the ACDC Challenge [1] in the absence of ground truth. Results show that our framework is accurate, fast and scalable, making it a viable option for quality control monitoring in clinical practice and large population studies.

3.1 Motivation

The current bottleneck towards the large scale use of learning-based pipelines in the clinics comes from the monitoring and maintenance of the deployed machine learning (ML) systems [47], assuring continuous high model performance and segmentation results. As shown in [1], despite the very high performances achieved, these methods may generate anatomically impossible results. In clinical practice and population studies, it is of utmost importance to constantly monitor a model's performance to determine when it degrades or fails, leading to poor quality results, as they may represent important risks. A system's continuous performance assessment and the detection of its degradation are challenging after deployment, due to the lack of a reference or ground truth. Therefore, translation of models into clinical practice requires the development of monitoring mechanisms to measure a

model's segmentation quality, in the absence of ground truth, that guarantee their safe use in clinical routine and studies.

Several factors could be cause of failures in both the acquisition process and the segmentation process of CMR images. Tarroni [48] asserts that the quality of a CMR image can easily be compromised, depending on the ability of the operator to correctly select the acquisition parameters in relation to the subject being scanned, the cooperation of the subject who must minimize movements during the process, and on some further circumstances out of the control of both the operator and the patient, such the presence of arrhythmias, the presence of bulk, the blood flow and the magnetic field inhomogeneities. The paper presents a fully automated quality control pipeline for CMR images, capable of detecting the scenarios described above to warn a human operator. When analysing 19265 short-axis (SA) cine stacks from the UKBB, such pipeline reports up to 14.2% with suboptimal coverage, up to 16% affected by noticeable inter-slice motion, up to 2.1% with an average end-diastolic cardiac image contrast below 30% of the dynamic range.

Insufficient image quality could cause a failure in any segmentation model and, in general, there is no way to determine a priori the behaviour of black box algorithms, such as neural networks, on new data. Nevertheless, when integrating automatic segmentation into a clinical setting it is of prime importance to be able to measure output quality, also in the absence of ground truth. Therefore, translation of models into practice requires the development of monitoring mechanisms to measure model performance, guaranteeing their safe use in clinical routine and studies. In a first attempt to assess performance in the absence of ground truth, Kohlberger [49] trained a model from segmentation errors measured against a ground truth, using a set of hand-crafted features, to predict overlap error and Dice Score Coefficient (DSC). To avoid the need of hand-crafted features, Robinson [50] proposed a supervised DL-based approach to predict Dice Similarity Coefficients (DSC) from estimated quality obtained via reverse testing strategy. This assessment is quick enough to allow an immediate intervention, such as a second scan in case the first is not analysable, both with or without the need of a human operator. Although, the results reported do not allow a direct one-to-one mapping to the reference DSC, but they need to be rounded on some threshold to correctly predict whether a segmentation is good or poor, without distinguishing between two segmentations of similar quality. More recently, Puyol-Antón [51] used a Bayesian neural network to measure a model performance by classifying its resulting segmentation as correct or incorrect.

The main drawback of the three methods mentioned in the paragraph above is that they require annotations reflecting a large set of quantitative (e.g. DSC) or qualitative (e.g. correct/incorrect) segmentation quality levels, which can be difficult to obtain. The Reverse Classification Accuracy (RCA) [52, 53] addressed this problem by using atlas label propagation. This registration-based method relies on the spatial overlap between predicted segmentations and reference atlases to measure the performance of a segmentation model on new data. The framework consists of a reverse classifier, trained from scratch on each new patient's scan, referring to the result of the original model as ground truth. Once trained, the

reverse classifier is tested on some reference images with available ground truth. This process works under the hypothesis that if the predicted segmentation is of good quality, then the reverse classifier will produce a good segmentation on at least one atlas image. However, training a new classifier to evaluate any new data has tight computational demands, and does not conform to a real-time application. Furthermore, it is possible that the atlas registration step fails. This is often the case for certain cardiac pathologies that introduce significant morphological deformations that the registration step is not able to recover [54]. In such cases, it is necessary to verify the results and manually fine-tune the registration step, limiting the method's scalability.

We present a novel learning framework to monitor the performance of cardiac image segmentation models in the absence of ground truth. Differently from previous learning-based approaches [49, 50, 51], we avoid the need of any type of annotations about the quality of a segmentation for training. Our approach also avoids the required spatial alignment between ground truth images and segmentations of RCA [53], thus circumventing image registration. The proposed monitoring framework relies on an anomaly detection setup with a convolutional autoencoder. Autoencoders can locate the cause of a low-quality verdict inside the segmentation itself. Reporting these occurrences to an expert, he could promptly focus on the precise part of the scan causing the failure in the segmentation model, speeding up eventual corrections. Since many relevant applications must rely on novelty detection protocols, the identification of anomalous structures in natural image data has been extensively covered in recent years. Machine learning systems seem to have difficulties to recognize if an image is similar or not to the images they previously observed, task efficiently performed by humans instead.

In 2019 Bergmann [55] conducted a thorough evaluation of current state-of-the-art unsupervised anomaly detection methods based on deep architectures. Among the many approaches suggested over the years, the most effective make use of Deep Convolutional Autoencoders (CAEs) [56], Generative Adversarial Networks (GANs) [57] or features appositaly extraced from pre-trained convolutional neural networks. In this work, CAEs are subjects of further experiments and readjustments for the identification of inconsistencies within segmentation masks.

3.2 Method

We first present our cardiac segmentation model monitoring framework, followed by a description of the selected architecture and training settings. Hereafter, we refer to 2D images and convolutional autoencoders, our backbone network. The extension to 3D images or other types of autoencoders is trivial.

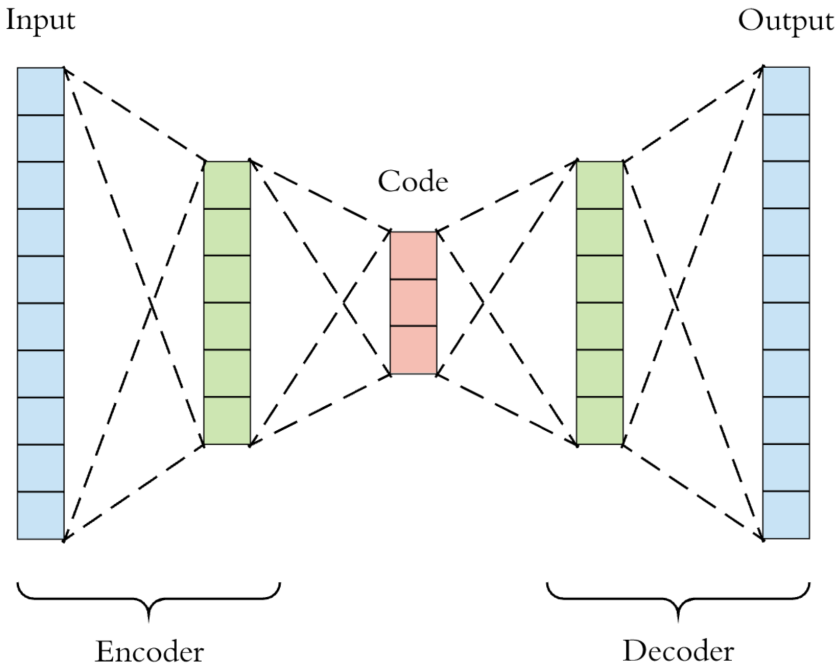

Figure 3.1. Architecture of an autoencoder.

3.2.1 Background

An autoencoder (Fig. 3.1) is an artificial neural network which aims to copy its input to its output, made of an encoding and a decoding part. Through the encoder, the input is projected into a latent representation, a code which encapsulates the information needed to the decoder for the reconstruction.

Autoencoders are designed to be unable to learn to copy perfectly: in this way, the model is forced to prioritize which aspects of the input should be copied, often learning useful properties of the data. The most common way to obtain useful features is to constrain the latent representations to lay in a lower dimensional space than the input space. An autoencoder with such property is called undercomplete. The learning process of an autoencoder consists of minimizing a loss function $\mathcal{L}_{\text{AE}}$ which retrieves larger penalties the more dissimilar the output is from the input. The main challenge when designing an autoencoder is the choice of the capacity for the encoder and the decoder, together with the choice of the latent size. There is not a strong mathematical theory to support these choices, and bad decisions lead to perform the copying task without extracting useful information.

3.2.2 Model Monitoring Framework

Let us denote $X \in \mathcal{C}^{H \times W}$ a segmentation mask of width W and height H, with $\mathcal{C}$ the set of possible label values. A Convolutional Autoencoder (CA) is trained to learn a function

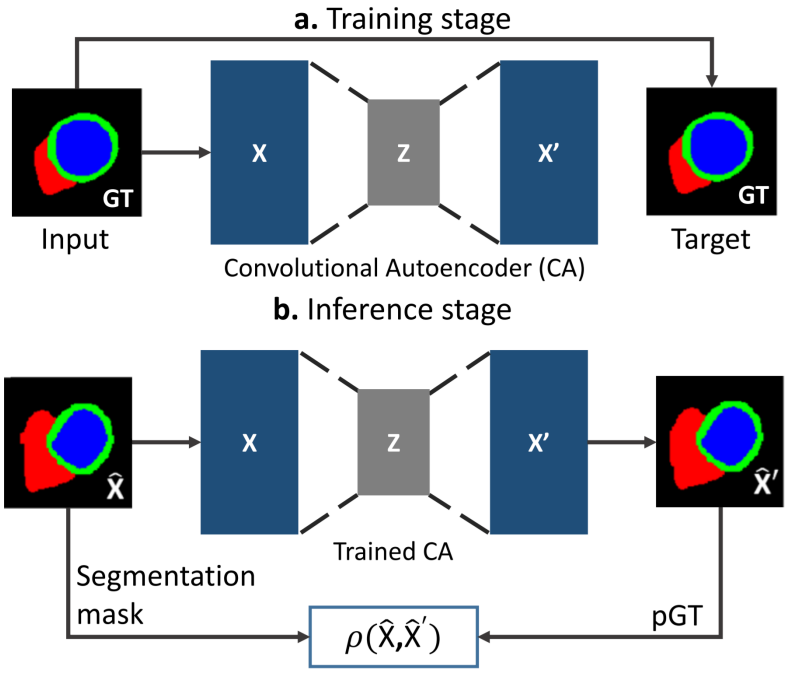

Figure 3.2. A convolutional autoencoder (CA) is trained with ground truth (GT) masks from a cardiac magnetic resonance (CMR) dataset. At inference, the CA reconstructs an input mask $\widehat{X}$, previously segmented by a given model. The reconstructed mask ($\widehat{X}'$) acts as a pseudo ground truth (pGT) to estimate a function ρ, a surrogate measure of the segmentation quality and the model performance.

$$f : \mathcal{C}^{H\times W} \rightarrow \mathcal{C}^{H\times W}, \tag{3.1}$$

with $X' = f(X) \approx X$, by minimizing a global dissimilarity measure between an input mask X and its reconstruction X'. In an anomaly detection setup, the CA is trained using normal samples, i.e., samples without defects. In our framework, the normal samples are the ground truth (GT) masks associated to the images used to train a segmentation model (Fig. 3.2a). The CA learns to reconstruct defect-free samples, i.e. the GT, through a bottleneck, the latent space Z.

At inference (Fig. 3.2b), the CA is used to obtain $\widehat{X}' = f(\widehat{X})$, where $\widehat{X}$ is a segmentation mask, generated by a cardiac segmentation model/method on unseen data, and $\widehat{X}'$ its reconstruction. Since the CA is trained with ground truth data, the quality of the reconstruction will be generally high for segmentation masks with similar characteristics than those in the ground truth. Poor segmentations, which the CA has not encountered at training, will instead lead to bad reconstructions ($\widehat{X}' \not\approx \widehat{X}$). Indeed, it will not be possible to find a proper representation of the input within Z, forged to exclusively capture intrinsic traits of correct segmentations from the GT. Autoencoder-based anomaly detection methods exploit the reconstruction error, i.e. $\|\widehat{X}' - \widehat{X}\|_2$, to quantify how anomalous is a sample [58].

We use this principle to establish a surrogate measure of the segmentation quality by quantifying a segmented mask and its reconstruction.

Let us so formalize the function $\rho(\widehat{X}, \widehat{X}')$, a surrogate measure of the segmentation quality of the mask in the absence of GT. In this context, we denote $\widehat{X}'$ a pseudo GT (pGT) since it acts as the reference to measure performance. We present two different scenarios for ρ. First, we propose

$$\rho_1 : \mathcal{C}^{H \times W} \rightarrow R, \tag{3.2}$$

which represents the most commonly setup in autoencoder-based anomaly detection, ρ_1 often being the L_2-norm. Due to the generic nature of ρ_1, better suited metrics for segmentation quality assessment can be used instead, such as the DSC or the Hausdorff Distance (HD). Secondly, we propose

$$\rho_2 : \mathcal{C}^{H \times W} \rightarrow R^{H \times W}. \tag{3.3}$$

This function generates a visual map of the inconsistencies between the two masks. We use as $\rho_2(\cdot)$ a pixel-wise XOR operation between the segmentation mask $\widehat{X}$ and the pGT.

These two types of measures can be used jointly for performance assessment and model monitoring. Measures obtained from ρ_1-type functions (Def. 3.2) can be paired with a threshold to flag poor segmentation results. The raised alert would then be used to take application-specific countermeasures as, for instance, a visual inspection of an inconsistency map generated by ρ_2-type functions (Def. 3.3).

3.2.3 Network Architecture

We use the CA architecture proposed in [59] as the backbone network (Fig. 3.3), introducing the following modifications. We use a latent space dimension to accommodate 100 features maps of size 4×4. A softmax activation function is added to the last layer to normalize the output to a probability distribution over predicted output classes, as well as batch-normalization and dropout to each hidden layer. We use the loss function:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}}(X, X') + \mathcal{L}_{\text{GD}}(X, X') \tag{3.4}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error loss and $\mathcal{L}_{\text{GD}}$ the generalized dice loss [60]. Trained over 500 epochs, for the first 10 epochs $\mathcal{L}_{\text{GD}}$ is computed leaving aside the background class to avoid the convergence to a dummy blank solution. The network weights are set using a He normal initializer. The Adam optimizer is initialized with learning rate 2×10^{-4} and a weight decay of 1×10^{-5}. After every epoch, the model is evaluated on the validation set. The weights retrieving the lowest $\mathcal{L}$ value are stored for testing.

Layer	Output Size	Parameters		
		Kernel	Stride	Padding
Input	256x256x4			
Block1	128x128x32	4x4	2	1
Block2	64x64x32	4x4	2	1
Block3	32x32x32	4x4	2	1
Block4	32x32x32	3x3	1	1
Block5	16x16x64	4x4	2	1
Block6	16x16x64	3x3	1	1
Block7	8x8x128	4x4	2	1
Block8	8x8x64	3x3	1	1
Block9	8x8x32	3x3	1	1
Conv2d	4x4x100	4x4	2	1

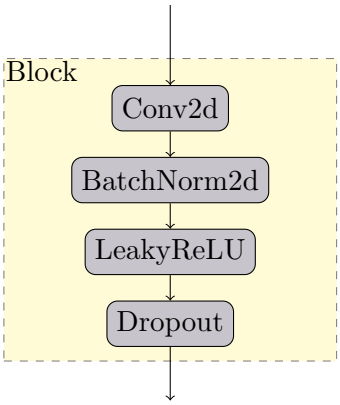

Figure 3.3. Architecture of the encoding module. The decoder is built reversing this structure, and replacing convolutions with transposed convolutions.

3.3 Experimental Setup

3.3.1 Data

We used data from the Automatic Cardiac Diagnosis Challenge (ACDC) [1]. The dataset consists of an annotated set with 100 short-axis cine magnetic resonance (MR) images, at end diastole (ED) and end systole (ES), with corresponding labels for the left ventricle (LV), right ventricle (RV) and myocardium (MYO). The set was split into training and validation subsets using a 80:20 ratio. The challenge also provides a testing set with 50 cases, with no ground truth publicly available. To have uniform image sizes, these were placed in the middle of a 256×256 black square. Those exceeding this size were center cropped.

3.3.2 Setup

We trained the monitoring framework using the ground truth masks from the ACDC training set and used it to assess the performance of five methods participating in

the ACDC Challenge [31, 28, 38, 35, 45] and an additional state-of-the-art cardiac segmentation model [46]. We trained five of these models [31, 28, 38, 35, 46] using the challenge's full training set (MR images and masks) and then segmented the ACDC test images. For the remaining method [45], we obtained the segmentation masks directly from the participating team. We gained access to the code of another model [61], but we were not able to reproduce results similar to those reported in the challenge [1]. Therefore, we discarded this model for the remaining experiments.

The segmentations from every method were fed to the monitoring framework. The resulting pGTs were used to compute ρ_1-type measures (Def. 3.2), the DSC and the HD, and a ρ_2-type measure, an inconsistency map (Def. 3.3). We also computed pseudo DSC/HD using the RCA [53]. The ACDC challenge platform estimates different performance measures (DSC, HD, and other clinical measures) on the testing set upon submission of the segmentation results. We uploaded the masks from every model to obtain real DSC and HD. To differentiate the real measures computed by the platform from our estimates, we denote the latter ones pDSC and pHD. In our experiments, we set pHD > 50 or pDSC < 0.5 to flag a segmentation as suspicious and pHD $=$ pDSC $= 0$ to raise an erroneous segmentation alert flag.

3.3.3 Implementation

We implemented our framework in Pytorch. All the cardiac segmentation models used the available implementations, except for [45], where we had the segmentation masks. The RCA was implemented following the guidelines in [53, 52] using a previously validated atlas propagation heart segmentation framework [22]. All experiments ran on Amazon Web Services with a Tesla T4 GPU. To encourage reproducibility, our code and experiments are publicly available[1].

3.4 Results

Figure 3.4 and Figure 3.5 present scatter plots of the real HD and DSC from the ACDC platform and the pHD and pDSC obtained with our framework and the RCA [53, 52]. We present results for LV, RV and MYO and report the Pearson correlation coefficient r.

We obtain $r(\text{HD}, \text{pHD}) = 0.950, 0.938$ and 0.971, $r(\text{DSC}, \text{pDSC}) = 0.715, 0.915$ and 0.769 for the LV, RV and MYO, which show a high correlation between real scores and our estimations. Our framework consistently outperforms RCA. The real and pseudo HDs show higher correlations than the DSC. This can be explained by

[1] https://github.com/robustml-eurecom/quality_control_CMR

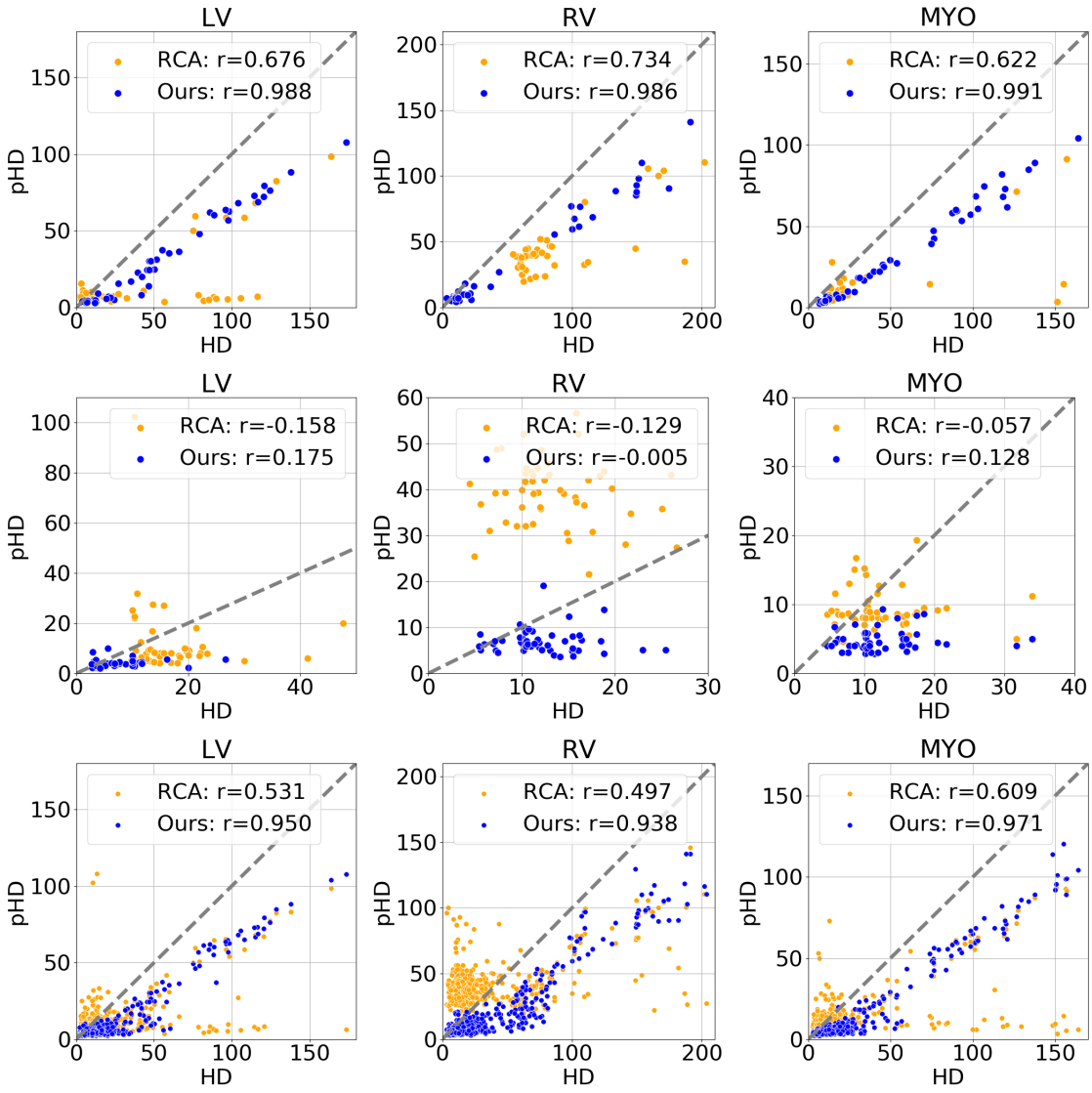

Figure 3.4. HD vs. pHD for our framework (CA) and RCA on LV (left), RV (center) and MYO (right). Models with highest (top) and lowest (middle) r, and all cases (bottom).

the higher sensitivity of the HD to segmentation errors. Instead, the DSC is robust to minor segmentation errors making accurate prediction of very high performing models is challenging, as the metric shows little variability.

In one case (Fig. 3.4 and Fig. 3.5 middle rows), our method fails to obtain pseudo measures that highly correlate to the real ones (and the same actually happens for RCA). This case presents low HDs/high DSCs, reflecting high quality segmentations. Although our framework predict low pHDs and high DSCs, it seems that an accurate prediction of very high performing models is challenging.

Table 3.1 and Table 3.2 simulate the ACDC Challenge results using real HD and pHD/DSC and pDSC for every model to determine if our framework is a reliable means to rank the performance of the different cardiac segmentation methods. We

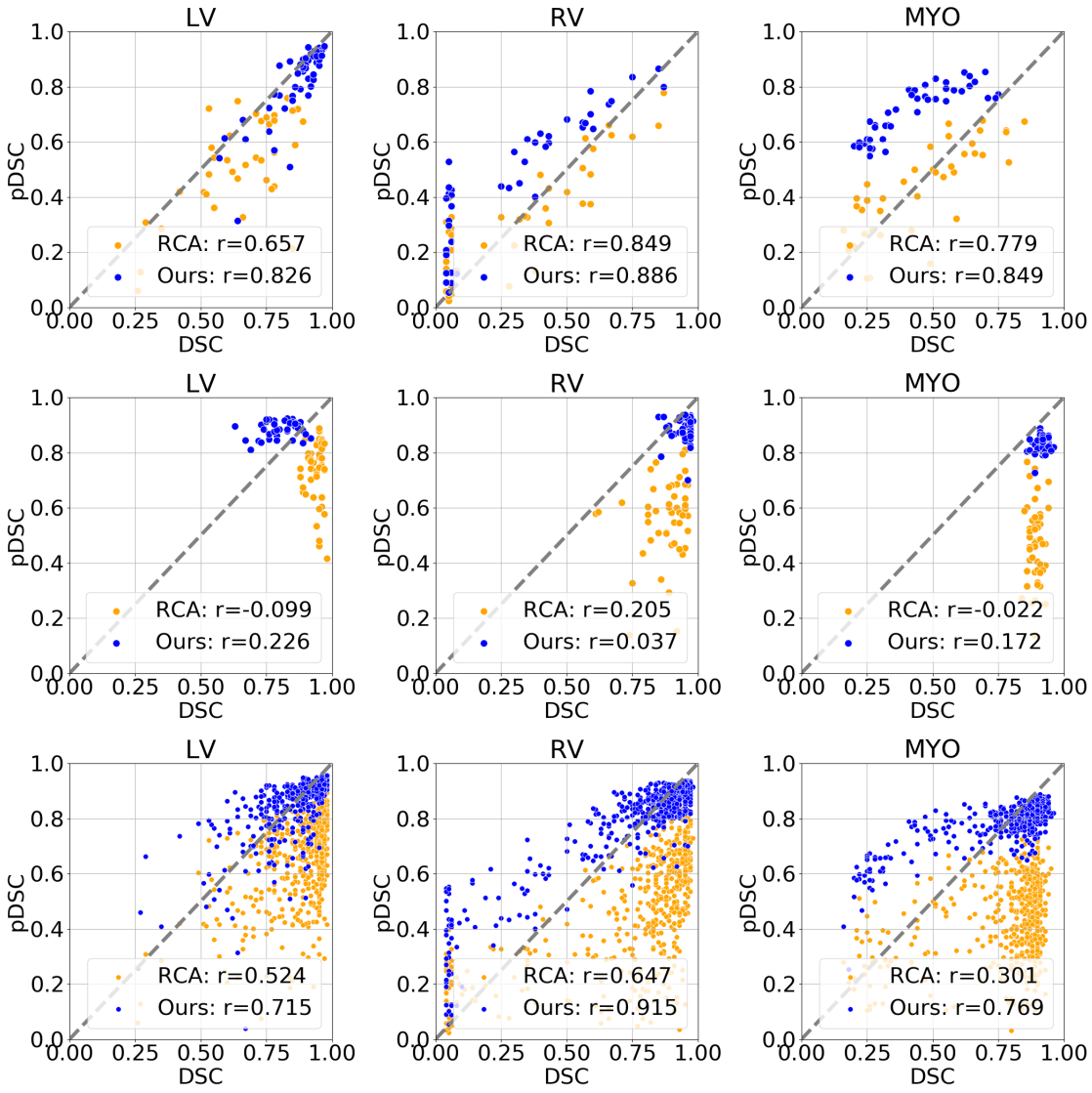

Figure 3.5. DSC vs. pDSC for our framework (CA) and RCA on LV (left), RV (center) and MYO (right). Models with highest (top) and lowest (middle) r, and all cases (bottom).

report results for LV, RV and MYO in ED and ES and compare them against the RCA. The ranking quality is assessed using the Spearman correlation coefficient r_s between the real and the pseudo measures, excluding one method [28] for which RCA failed. In the case of HD, which indeed shows a higher r coefficient (Fig. 3.4), we were able to reproduce the real ranking (r_s=1.0) for five out of six cases. In the remaining cases, there are only small differences between the real and our pseudo ranking, with the exception of DSC-ED-MYO (r_s=0.3) and DSC-ES-LV (r_s=0.1). This shows that our framework is a reliable mean for method ranking.

Through the use of alert flags we were able to detect 16 cases for which the challenge platform reported `NaN` values that indicate errors in the submitted results. Fifteen cases were flagged as erroneous (pHD = pDSC = 0) and one as suspicious (pHD > 50) by our framework. Fig. 3.6 shows the inconsistency maps of two cases.

Table 3.1. ACDC Challenge simulation reporting real HD (GT), pHD using the proposed framework (Ours) and RCA, and the Spearman correlation coefficient r_s between the the real and the pseudo measures for 6 models in ED and ES. pHD using RCA for [28] were excluded due to failures in the registration step.

ED									
	LV			RV			MYO		
Model	GT	**Ours**	RCA	GT	Ours	RCA	GT	**Ours**	RCA
Bai [46]	39.01	23.38	15.55	50.21	31.82	56.22	47.10	28.46	20.42
Baumgartner [31]	7.14	3.87	9.30	14.00	7.72	37.63	9.49	4.43	10.52
Isensee [28]	7.01	3.88	-	11.40	7.82	-	8.44	4.38	-
Khened [38]	16.81	6.39	10.58	13.25	6.87	39.01	16.09	6.08	11.22
Tziritas [45]	8.90	4.69	8.92	21.02	9.86	41.10	12.59	4.58	10.65
Yang [35]	16.95	5.29	12.96	86.08	47.24	44.75	31.93	16.39	15.12
r_s	-	0.90	0.90	-	1.00	0.80	-	1.00	1.00
ES									
	LV			RV			MYO		
Model	GT	**Ours**	RCA	GT	**Ours**	RCA	GT	**Ours**	RCA
Bai [46]	50.53	29.56	20.01	52.73	31.40	53.68	52.72	31.05	26.60
Baumgartner [31]	10.51	4.41	9.56	16.32	7.10	35.50	12.47	4.77	9.33
Isensee [28]	7.97	4.07	-	12.07	6.99	-	7.95	4.27	-
Khened [38]	20.14	6.96	11.72	14.71	7.07	35.65	16.77	6.03	10.36
Tziritas [45]	11.57	5.00	10.46	25.70	9.61	36.51	14.78	5.59	10.60
Yang [35]	19.13	6.11	11.78	80.42	33.21	40.68	32.54	16.98	13.68
r_s	-	1.00	0.90	-	1.00	0.80	-	1.00	0.90

Table 3.2. ACDC Challenge simulation reporting real DSC (GT), pDSC using the proposed framework (Ours) and RCA, and the Spearman correlation coefficient r_s between the the real and the pseudo measures for 6 models in ED and ES. pDSC using RCA for [28] were excluded due to failures in the registration step.

ED									
	LV			RV			MYO		
Model	GT	**Ours**	RCA	GT	Ours	RCA	GT	**Ours**	RCA
Bai [46]	0.96	0.93	0.80	0.94	0.88	0.67	0.89	0.80	0.45
Baumgartner [31]	0.96	0.93	0.73	0.93	0.89	0.61	0.88	0.81	0.38
Isensee [28]	0.97	0.93	-	0.95	0.89	-	0.90	0.81	-
Khened [38]	0.94	0.92	0.74	0.88	0.88	0.56	0.85	0.82	0.40
Tziritas [45]	0.95	0.92	0.74	0.86	0.86	0.57	0.79	0.80	0.43
Yang [35]	0.81	0.89	0.65	0.31	0.48	0.33	0.43	0.71	0.34
r_s	-	0.80	0.70	-	0.90	0.90	-	0.30	0.60
ES									
	LV			RV			MYO		
Model	GT	**Ours**	RCA	GT	**Ours**	RCA	GT	**Ours**	RCA
Bai [46]	0.85	0.87	0.67	0.86	0.84	0.53	0.86	0.82	0.54
Baumgartner [31]	0.89	0.84	0.60	0.86	0.86	0.43	0.89	0.82	0.48
Isensee [28]	0.93	0.85	-	0.90	0.85	-	0.92	0.83	-
Khened [38]	0.86	0.85	0.59	0.83	0.83	0.37	0.88	0.82	0.49
Tziritas [45]	0.87	0.81	0.56	0.74	0.80	0.37	0.80	0.80	0.44
Yang [35]	0.65	0.72	0.52	0.18	0.45	0.23	0.46	0.66	0.43
r_s	-	0.10	0.30	-	0.90	0.90	-	1.00	0.60

The top-most shows a segmentation flagged as erroneous, where the inconsistency map confirms that the LV has not been segmented. The bottom-most presents a segmentation flagged as suspicious, where the LV pHD is high (pHD = 104.93), although the pDSC = 0.814 is within normal range. The inconsistency map confirms the clear segmentation error.

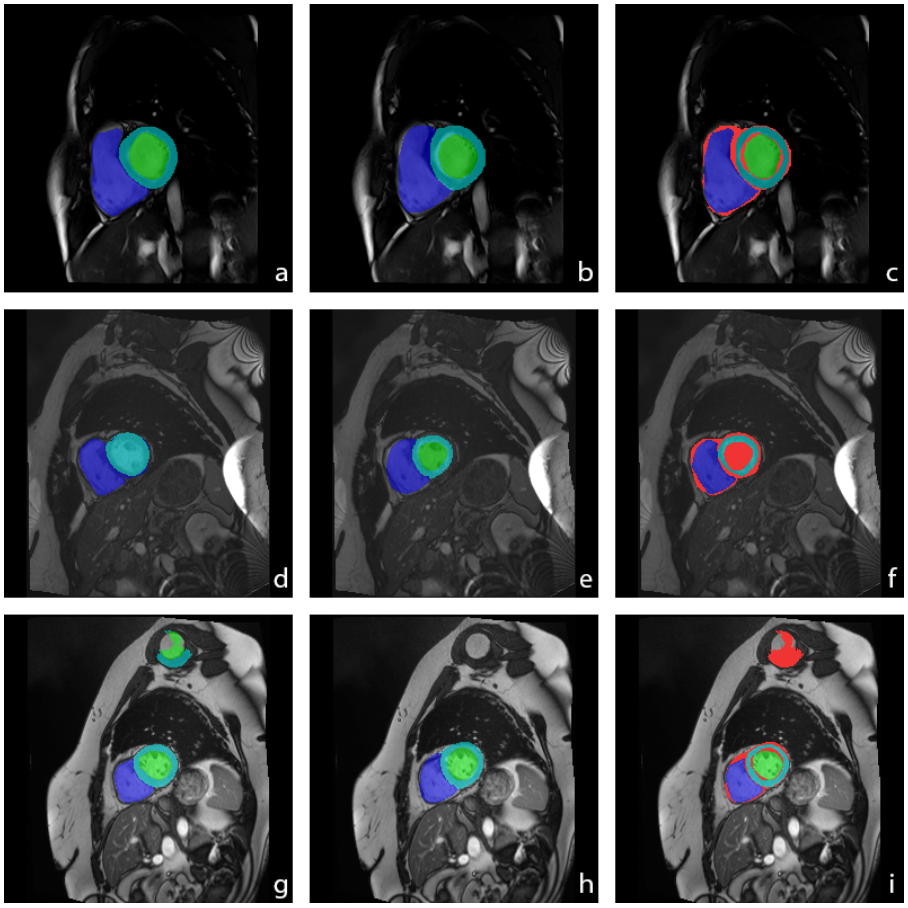

Figure 3.6. Segmentations (a, d and g), along with pGTs (b, e and h), and inconsistency maps (c, f and i). Case a shows a successful segmentation, which is indeed not flagged by our framework. Case d is flagged as erroneous with pHD = pDSC = 0 in the LV; case g is flagged as suspicious with pHD = 104.93 for the LV (pDSC = 0.814). The inconsistency maps confirm the segmentation errors.

3.5 Conclusions

We presented a novel learning framework to monitor the performance of cardiac image segmentation models in the absence of ground truth. Our framework addresses the limitations of previous learning-based approaches thanks to its formulation under an anomaly detection paradigm, which allows training without the need of quality scores labels. The reported results show a good correlation between real performance measures and those estimated with the pGT, making it a reliable alternative when there is no reference to assess a model.

Compared with state-of-the-art RCA, our method avoids the use of image registration which makes it more robust, scalable and considerably faster ($\approx$ 20 min RCA vs. $\approx$ 0.2 s ours, per case). CAs allow for fast inference which conforms to real-time use, thus permitting a quick quality assignment, for example, in a clinical setting. All these characteristics make the proposed framework a viable option for quality control and system monitoring in clinical setups and population studies.

We are proud to specify that the writing of this chapter led to the submission of a paper at FIMH 2021.

Chapter 4

Semi-supervised segmentation for improved cross-domain generalisation

In this chapter, we introduce a semi-supervised learning framework that builds upon the results achieved by our quality assessment model in Chapter 3. We train again a convolutional autoencoder to learn the variability of ground truth data in a source domain. Under the assumption that the label space is consistent across domains, the quality control (QC) module serves as a proxy of the segmentation model's performance when tested in unlabelled data from a target domain. This information is used as feedback to refine the training of the segmentation model, which learns from both labelled and unlabelled data and adapts to the target. We evaluated our method on a public multi -centre, -vendor and -disease cardiac MR image segmentation dataset [62]. We show that by using a single labeled source domain along with unlabelled data from the target domain, we increase the cross-domain generalisation of the segmentation model as measured by the Dice score coefficient.

4.1 Motivation

Deep learning methods still suffer a severe limitation: they fail to generalise to a domain different from the one of the training set [46], e.g., a different scanner, acquisition protocol or population demographics. In order to fulfill the task of CMR segmentation in a traditional supervised manner, data scientists need manually delineated cardiac images to train their models. Labeling data from the unseen domain to then re-train the original model, although straightforward, is expensive, labour-intensive and not scalable to clinical scenarios. The collection of a significant amount of data requires a major effort by radiologists, which are delegated to manually draw contours of the heart sections, an analysis that typically takes a trained expert around 20 minutes per subject. For this reason, it is hardly viable

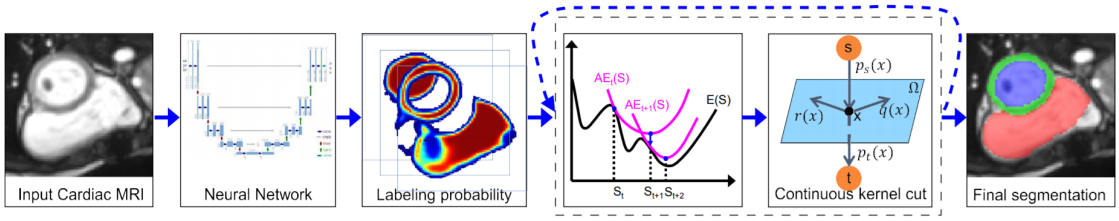

Figure 4.1. Segmentation workflow developed by Guo [67].

to include in an annotated training set enough variability to faithfully represent different demographics, protocols and scanners.

An example of the generalisation problem pointed out in the previous paragraph can be found in Bai's work [46]. The authors combined cardiac images from the ACDC dataset [1] and the UK Biobank dataset [2], counting respectively 100 and 5008 patients. They showed that once trained on one dataset, their model did not generalise well when tested on the other unless performing fine-tuning to force the adaption of the model to the new data. Since fine-tuning requires resuming the training process every time the protocol or the scanner in use changes, this is not a practicable solution. The reasons behind such a disappointing result have to be sought in the major differences between the two datasets under study. Besides the fact that images were taken with different scanners and protocols, the UK Biobank dataset is more homogeneous, with a preponderance of healthy cases. On the contrary, a wide proportion of the ACDC dataset consists of pathological cases, some absent in the UK Biobank cohort due to the variety of clinical patterns.

Model generalisation is an active topic of research beyond medical imaging [63, 64], and, over the past years, multiple works have explored alternatives [65, 66, 67, 68, 69] to improve the generalisation capacity of CMR segmentation models. U-Net based models count millions of parameters, so previous methods have tackled the problem by reducing the model's complexity using regularisation [67] or optimised network architectures that reduce the number of parameters [68]. Khened [68] presented a DenseNet-based FCN architecture, whose optimized connectivity pattern leads to parameter efficiency and, consequently, to better generalisation. Guo [67] developed a model integrating CNN, continuous kernel cut, and bound optimization in a unified max-flow framework which was demonstrated to improve model generalisability. [Fig. 4.1] Although these techniques are very effective at reducing overfitting when the training sets are small, there is no guarantee they can mitigate poor cross-domain generalisation.

An alternative to improve model generalisation is to increase the amount of available training data. Data augmentation has been explored to enlarge the training set by simulating various possible data distributions across different domains, applying geometrical operations to the source training data. Chen [65] proposed a training pipeline for CNN-based cardiac segmentation methods revealing simple but effective data normalization and augmentation strategies that improve generalisability. The technique, however, has shown to be less performing with cross-domain data

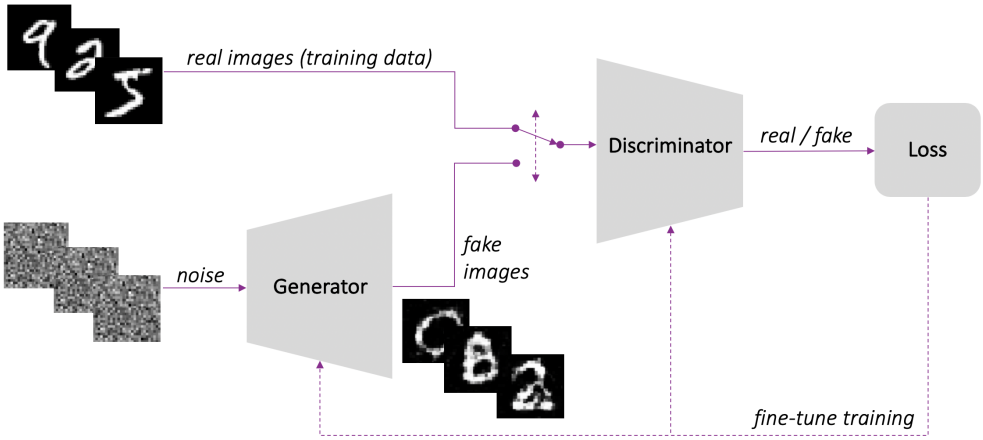

Figure 4.2. The architecture of a GAN, a generative model capable of drawing samples with the same statistics as in a given dataset. Its functioning depends on the outcome of a game played by two neural networks, the generator and the discriminator.

than with the intra-domain ones [65]. Domain adaptation techniques propose to enlarge the training set by combining labeled source domain data with target domain one. Depending on how the target domain data is exploited, these methods can be unsupervised [69, 70] or semi-supervised [66, 71].

Nie [71] developed an attention-based semi-supervised deep learning framework integrating a fully convolutional confidence network to adversarially train a pelvic organ segmentation model. Such setup was originally introduced by Hung [72] in a more general context. The framework substitutes the generator and the CNN-based discriminator of a Generative Adversarial Network (GAN) [57] [Fig. 4.2] with a segmentation network and an FCN-based discriminator respectively. Instead of a binary output, an FCN-based discriminator retrieves a confidence map of the same width and height as the input. Each value of the confidence map represents the probability of being peculiar to the ground truth domain or the segmentation output domain at a local level. This information points out the trustworthy regions in the label maps output by the segmentation network on unlabelled data, which are therefore integrated into the backpropagation learning process.

Chen [70] presented an unsupervised domain adaptation framework, named as Synergistic Image and Feature Alignment (SIFA), to adapt a segmentation network to an unlabeled target domain. The proposed model conducts synergistic alignment of domains from both image and feature perspectives. The authors leveraged adversarial learning and deeply supervised mechanism to simultaneously transform the appearance of images across domains and enhance domain-invariance of the extracted features [Fig. 4.3].

For both the methods described above [71, 70], domain adaptation approaches rely on adversarial training, which, however, represents their main limitation. Indeed, it has been well-established that adversarial training is difficult and prone to instabilities due to problems such as mode collapse and non-convergence [73].

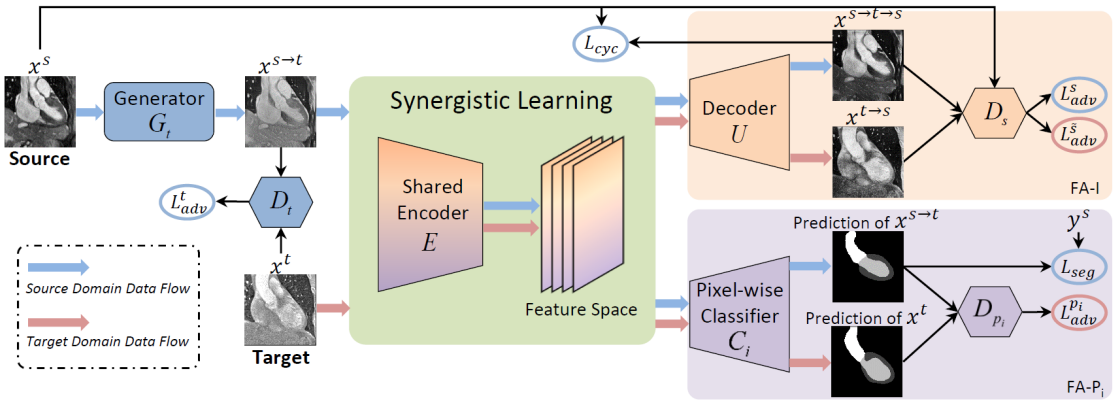

Figure 4.3. Architecture developed by Chen [70].

In this chapter, we propose to exploit QC information to assist a segmentation model to learn from unlabelled data coming from a new domain. Under the assumption that the label space is consistent across domains, we train a QC assessment module to learn the variability of ground truth data in a source domain. At inference, the QC module provides estimates of a segmentation model's performance in unseen data from the target domain. We use these quality measurements as feedback to refine the training of a segmentation model, which was previously trained using source domain data. To the best of our knowledge, although segmentation quality measurements are a proxy of a model's generalisation capabilities, this information has not been explicitly used to improve cross-domain generalisation before. Formulated as a semi-supervised process, our training framework avoids the limitations of data augmentation strategies. Furthermore, our QC module allows us to avoid the adversarial setup of domain adaptation techniques, thus leading to improved robustness and stability.

4.2 Method

Let us denote $X \subset R^{H \times W}$ the input image domain, and $Y \subset \mathcal{C}^{H \times W}$ the domain of the corresponding segmentation masks with width W and height H, where $\mathcal{C}$ is the set of possible label values. Our goal is to train a segmenter network $\mathcal{M}$ using $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, a training set from a given source domain s, consisting of n_s source images $\mathbf{x}_i^s \in X$ and the corresponding segmentation masks $\mathbf{y}_i^s \in Y$. In addition to $\mathcal{D}_s$, we have access to $\mathcal{D}_t = \{(\mathbf{x}_i^t)\}_{i=1}^{n_t}$ an unlabelled set in a target domain t, which is different from s.

To achieve cross-domain generalisation, i.e. preventing the performance drop of $\mathcal{M}$ when tested in t, we propose a learning framework, in which the generalisation of the segmenter $\mathcal{M}$ is enabled by a QC assessment module. Assuming that the space Y is consistent across domains, the QC module consists of a reconstruction

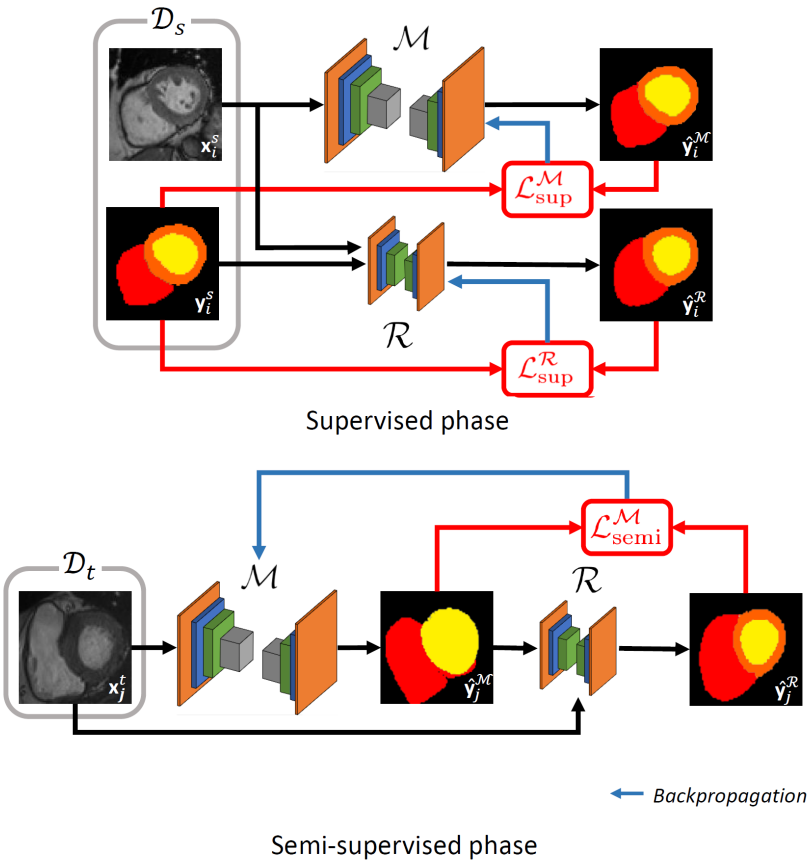

Figure 4.4. Illustration of our method. During the supervised phase, both the segmenter $\mathcal{M}$ and the image reconstructor $\mathcal{R}$ are trained independently using $\mathcal{D}_s$. During the semi-supervised phase, the pre-trained segmenter $\mathcal{M}$ is used to segment unlabelled images from $\mathcal{D}_t$ that are then fed to $\mathcal{R}$. The difference between the reconstruction $\widehat{\mathbf{y}}^R$ and the model's segmentation $\widehat{\mathbf{y}}^M$ is backpropagated to update $\mathcal{M}$.

network $\mathcal{R}$, which quantifies if a segmentation mask provided by the segmenter in the domain t is compatible with the general variability of ground truth data learned from domain s.

The framework is trained in two phases. First, $\mathcal{M}$ and $\mathcal{R}$ are individually trained on $\mathcal{D}_s$ (*supervised phase*). Second, using $\mathcal{D}_t$, $\mathcal{M}$ is updated to achieve realistic segmentations in the target domain t according to the QC feedback given by $\mathcal{R}$ (*semi-supervised phase*). Figure 4.4 illustrates the proposed framework scheme.

4.2.1 The segmenter $\mathcal{M}$

The segmenter network $\mathcal{M}$ learns a function $f_{\mathcal{M}} : X \to Y$, which is then used to predict a segmentation mask $\widehat{\mathbf{y}}^M = f_{\mathcal{M}}(\mathbf{x})$. In this sense, it is a standard segmentation network trained in a supervised setting, where the training and validation

sets come from the same source domain s. It should be noted that the functioning of the framework is not conceived to depend on a specific segmenter architecture, for which several options are available in the literature tailored for cardiac image segmentation [74].

4.2.2 The image reconstructor $\mathcal{R}$

In line with Chapter 3, we use an anomaly detection setup for QC of image segmentations. In this context, the image reconstructor $\mathcal{R}$ is trained to learn a function $f_R : X \times Y \to Y$, with

$$\widehat{\mathbf{y}}^R = f_R(\mathbf{x}, \mathbf{y}) \approx \mathbf{y}. \tag{4.1}$$

In an anomaly detection setup, $\mathcal{R}$ is trained using normal samples, *i.e.* samples without defects. In our framework, the normal samples are the ground truth masks coming from $\mathcal{D}_s$. The reconstructor learns to recreate defect-free samples, *i.e.* the ground truth, through a bottleneck.

Under the assumption that the space Y is consistent across domains, once trained, $\mathcal{R}$ is used to obtain $\widehat{\mathbf{y}}^R = f_R(\mathbf{x}, \widehat{\mathbf{y}}^M)$, where $\widehat{\mathbf{y}}^M$ is a segmentation mask, generated by the cardiac segmenter $\mathcal{M}$ on unseen data, and $\widehat{\mathbf{y}}^R$ its reconstruction. Since $\mathcal{R}$ is trained with ground truth data, the quality of the reconstruction will be generally high for segmentation masks with similar characteristics than those in the ground truth. Poor segmentations, which $\mathcal{R}$ has not encountered at training, will instead lead to bad reconstructions ($\widehat{\mathbf{y}}^R \not\approx \widehat{\mathbf{y}}^M$).

Autoencoder-based anomaly detection methods measure the degree of (dis-) similarity between the input ($\widehat{\mathbf{y}}^M$) and its reconstruction ($\mathbf{y}^R$) and use this information as a surrogate measure for QC of the segmentation. We use this principle to measure the performance of $\mathcal{M}$ in the target domain t and then backpropagate it to refine $\mathcal{M}$.

Our reconstructor is implemented as a modified convolutional autoencoder. Differently from the original conception [59], the autoencoder network receives as input both the segmentation mask to reconstruct $\mathbf{y}_i$ and its associated image $\mathbf{x}_i$. The output remains only the reconstructed segmentation mask (Eq. 4.1).

4.2.3 Two-phase training

In this section we look at the details of the two-phase training process, which is summarised in Algorithm 1.

Algorithm 1 Two-phase training algorithm

Require: Datasets $\mathcal{D}_s$, $\mathcal{D}_t$, threshold λ
Ensure: Segmenter $\mathcal{M}$

Initialize $\mathcal{M}, \mathcal{R}$
repeat
  Sample $(\mathbf{x}_i^s, \mathbf{y}_i^s)$ from $\mathcal{D}_s$
  Forward pass:
  $\widehat{\mathbf{y}}_i^M \leftarrow f_M(\mathbf{x}_i^s)$
  Drop-connect($\mathcal{R}$)
  $\widehat{\mathbf{y}}_i^{\mathcal{R}} \leftarrow f_R(\mathbf{x}_i^s, \mathbf{y}_i^s)$
  Estimate losses:
  $\mathcal{L}_{\text{sup}}^M \leftarrow \mathcal{L}(\widehat{\mathbf{y}}_i^M, \mathbf{y}_i^s)$
  $\mathcal{L}_{\text{sup}}^R \leftarrow \mathcal{L}(\widehat{\mathbf{y}}_i^{\mathcal{R}}, \mathbf{y}_i^s)$
  Back-propagate $\mathcal{L}_{\text{sup}}^M$, $\mathcal{L}_{\text{sup}}^{\mathcal{R}}$
  Update model parameters $\mathcal{M}, \mathcal{R}$
until stopping criteria met
repeat
  Sample $(\mathbf{x}_i^s, \mathbf{y}_i^s)$ from $\mathcal{D}_s$
  Sample $\mathbf{x}_j^t$ from $\mathcal{D}_t$
  Forward pass:
  $\widehat{\mathbf{y}}_i^M \leftarrow f_M(\mathbf{x}_i^s)$
  $\widehat{\mathbf{y}}_j^M \leftarrow f_M(\mathbf{x}_j^t)$
  Reconstruct:
  $\widehat{\mathbf{y}}_j^R \leftarrow f_R(\mathbf{x}_j^t, \widehat{\mathbf{y}}_j^M)$
  Estimate losses:
  $\mathcal{L}_{\text{sup}}^M \leftarrow \mathcal{L}(\widehat{\mathbf{y}}_i^M, \mathbf{y}_i^s)$
  $\mathcal{L}_{\text{semi}}^M \leftarrow \mathcal{L}(\widehat{\mathbf{y}}_j^M, \widehat{\mathbf{y}}_j^{\mathcal{R}})$
  Back-propagate $\mathcal{L}_{\text{sup}}^M, \lambda\mathcal{L}_{\text{semi}}^M$
  Update model parameters $\mathcal{M}$
until stopping criteria met

Step 1: Supervised phase.

During the supervised phase, $\mathcal{M}$ and $\mathcal{R}$ are trained individually on the labelled set $\mathcal{D}_s$. $\mathcal{M}$ is trained to minimise a loss function measuring the dissimilarity between the ground truth masks $\{\mathbf{y^s}\}$ and the model's prediction $\widehat{\mathbf{y}}^M$:

$$\mathcal{L}_{\text{SUP}}^M = \mathcal{L}_{\text{GD}}(\widehat{\mathbf{y}}^M, \mathbf{y^s}), \tag{4.2}$$

with $\mathcal{L}_{\text{GD}}$ generalised dice loss [60]. The reconstructor $\mathcal{R}$ uses the loss

$$\mathcal{L}_{\text{SUP}}^R = \mathcal{L}_{\text{MSE}}(\widehat{\mathbf{y}}^R, \mathbf{y}^s) + \mathcal{L}_{\text{GD}}(\widehat{\mathbf{y}}^R, \mathbf{y}^s), \tag{4.3}$$

where $\mathcal{L}_{\text{MSE}}$ is the mean squared error loss and $\mathcal{L}_{\text{GD}}$ the generalised dice loss.

To ensure robustness, we apply drop-connect [75] to the input layer of the image $\mathbf{x}$ channel by multiplying the weights by a Bernoulli distributed variable u:

$$u \sim \mathcal{B}(p)$$

with probability p of being 1. We set $p = 0.5$.

Step 2: Semi-supervised phase.

Trained in the supervised phase, $\mathcal{R}$ is no longer subject of training. The semi-supervised training phase seeks to refine $\mathcal{M}$ by back-propagating information from the reconstructor $\mathcal{R}$, helping $\mathcal{M}$ to generalise better. The training alternates labelled and unlabelled data, respectively from $\mathcal{D}_s$ and $\mathcal{D}_u$.

During the forward pass, $\mathcal{M}$ predicts a segmentation mask $\widehat{\mathbf{y}}^M$ using data from either $\mathcal{D}_s$ or $\mathcal{D}_t$. If the input sample has been drawn from $\mathcal{D}_s$, a loss is computed as in Eq. 4.2. When the input sample comes from $\mathcal{D}_t$, as there is no ground truth data, $\mathcal{R}$ is used to obtain a surrogate measure of the generalisation capabilities of $\mathcal{M}$. The reconstructor is fed with $\widehat{\mathbf{y}}^M$ and estimates $\widehat{\mathbf{y}}^R$ (Eq. 4.1). The similarity of the two segmentation masks is measured through the loss:

$$\mathcal{L}_{\text{SEMI}} = \mathcal{L}_{\text{WGD}}(\widehat{\mathbf{y}}^M, \widehat{\mathbf{y}}^R), \tag{4.4}$$

with $\mathcal{L}_{\text{WGD}}$ the weighted generalised dice loss. Finally, both losses are combined into a total loss:

$$\mathcal{L}_{\text{TOTAL}} = \mathcal{L}^M_{\text{SUP}} + \lambda \mathcal{L}_{\text{SEMI}}, \tag{4.5}$$

with λ is a scaling hyper-parameter factor which reflects the reliability of the reconstruction. The loss is backpropagated to refine the training of $\mathcal{M}$.

4.2.4 Implementation.

We adapted a state-of-the-art cardiac segmentation network [28][1], the winner of the ACDC Challenge [1], as the segmenter network. Originally written in Theano, we re-implemented it in PyTorch. For the sake of training speed-up, we use only the 2D submodel from the original ensemble method. The model was trained using an Adam optimiser, with an initial learning rate of 5e−4, which was decayed by 0.985 at each epoch, and a weight decay of 1e−5. The network weights were initialised using a He normal initialiser [76]. For the reconstructor, we extended

[1]https://github.com/MIC-DKFZ/ACDC2017

the convolutional autoencoder [59] implementation in Chapter 3 to account for the concatenated input consisting of segmentation masks and images. The value of λ (Eq. 4.5) was selected via hyperparameter tuning among 1e-{1,2,3}. All code was written in PyTorch, and run on Amazon Web Services with a Testa T4 GPU.

4.3 Experiments and Results

4.3.1 Experimental Setup

Data. We used data from the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms) [62]. The dataset consists of 345 CMR images, at end diastole (ED) and end systole (ES), with corresponding labels for the left ventricle (LV), right ventricle (RV) and myocardium (MYO). Data were collected from five different centres in three different countries, which used four different MR scanner vendors. The set was split into five folds, one per centre, and further divided into training and validation subsets using a 80:20 ratio. CMR scans were standardised to have zero mean and standard deviation of 1. To have uniform image sizes, these were placed in the middle of a 352×352 black square. Those exceeding this size were center cropped.

Setup. We trained $\mathcal{M}$ and $\mathcal{R}$ using data from one vendor at a time. We split the data at the subject-scan level, using 70, 10, 20 % cases for training, validation and testing. The data from the vendors not used in the supervised phase was used in the semi-supervised phase. The segmenter was tested in the reserved test splits from all the vendors. We used the Dice score and the Hausdorff distance to evaluate the segmentation results.

4.3.2 Results

Benchmark. We compare our method with several alternatives. Specifically, we evaluate: i) the segmenter $\mathcal{M}$, trained with $\mathcal{D}_s$, as the reference baseline (REF), ii) a state-of-the-art data augmentation method for cardiac segmentation [65] (DA) in combination with the REF baseline and iii) ADS-Net a state-of-the-art semi-supervised method for image segmentation [71]. We implemented the data augmentation method following the guidelines in [65]. For ADS-Net, we used the available code as starting point. Hyper-parameter tuning was performed following the authors' guidelines.

Table 4.1 and Table 4.2 summarise the quantitative results using the Dice score in ED and ES respectively, where we also include the use of DA with our framework. Fig. 4.5 presents qualitative results. The results show model brings clear

Figure 4.5. Qualitative results of benchmark performance. From the left to the right: GT, REF, REF+DA, ASDNet, ours.

benefits, improving the cross-domain generalisation of $\mathcal{M}$ with respect to the baseline model (REF). We see that data augmentation brings some benefits although less significant than the ones reported by our method. This is consistent with previous results reported in the literature [65]. Regarding ASD-Net, we highlight the difficulties in the training process. The loss of this network is the weighted sum of three losses: a supervised one, a semi-supervised one, and an adversarial one. The adversarial term is due to the presence of a discriminator and a generator that fight one another. This process is, in general, hard to balance and it can cause the model to diverge. In addition, the ASD-Net contains several hyperparameters, whose tuning is required. Small changes in the learning rates of the generator and the discriminator can have relevant effects on the general performance. In addition, being the total loss a weighted sum of three different terms, it is necessary to find the correct values for the three weights.

Table 4.1. Quantitative evaluations reporting mean Dice score and standard deviation (in parenthesis) for ED.

	A			B			C			D		
A	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	**0.93**	0.81	**0.88**	**0.93**	0.79	0.88	0.83	0.62	0.70	0.89	0.72	0.83
	(0.06)	(0.11)	(0.06)	(0.03)	(0.07)	(0.07)	(0.18)	(0.18)	(0.27)	(0.07)	(0.07)	(0.10)
REF+ DA	0.92	0.82	0.87	0.92	0.78	**0.90**	0.86	0.67	**0.85**	**0.91**	**0.77**	**0.89**
	(0.07)	(0.07)	(0.10)	(0.04)	(0.09)	(0.05)	(0.07)	(0.12)	(0.08)	(0.04)	(0.06)	(0.07)
ASDNet	0.90	0.78	0.86	0.86	0.69	0.86	0.79	0.56	0.66	0.82	0.60	0.86
	(0.05)	(0.06)	(0.07)	(0.07)	(0.06)	(0.05)	(0.10)	(0.12)	(0.32)	(0.10)	(0.09)	(0.07)
Ours	0.93	**0.83**	0.87	0.92	**0.79**	0.87	**0.88**	**0.71**	0.70	0.87	0.74	0.86
	(0.04)	(0.05)	(0.06)	(0.04)	(0.06)	(0.05)	(0.09)	(0.09)	(0.27)	(0.04)	(0.07)	(0.06)
Ours + DA	0.93	0.81	0.87	0.89	0.72	0.81	0.86	0.66	0.65	0.87	0.69	0.83
	(0.04)	(0.05)	(0.06)	(0.04)	(0.06)	(0.08)	(0.05)	(0.09)	(0.24)	(0.03)	(0.07)	(0.08)
	A			B			C			D		
B	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.54	0.38	0.44	0.93	0.81	0.87	0.81	0.67	0.71	0.88	0.73	0.82
	(0.30)	(0.26)	(0.33)	(0.09)	(0.12)	(0.14)	(0.20)	(0.18)	(0.30)	(0.05)	(0.12)	(0.13)
REF + DA	0.36	0.29	0.22	0.93	0.80	0.85	**0.88**	**0.73**	**0.84**	0.87	0.66	0.66
	(0.33)	(0.28)	(0.32)	(0.06)	(0.11)	(0.15)	(0.06)	(0.10)	(0.10)	(0.09)	(0.21)	(0.26)
ASDNet	0.74	0.56	0.54	0.93	**0.82**	0.88	0.86	0.72	0.68	**0.89**	0.74	0.82
	(0.27)	(0.27)	(0.32)	(0.05)	(0.09)	(0.12)	(0.10)	(0.13)	(0.32)	(0.04)	(0.09)	(0.12)
Ours	0.74	0.54	0.60	0.91	0.79	0.84	0.84	0.72	0.62	0.85	0.71	0.77
	(0.25)	(0.23)	(0.26)	(0.05)	(0.08)	(0.05)	(0.11)	(0.09)	(0.26)	(0.06)	(0.08)	(0.08)
Ours + DA	**0.81**	**0.66**	**0.68**	**0.93**	0.81	**0.90**	0.82	0.68	0.71	0.88	**0.74**	**0.85**
	(0.21)	(0.19)	(0.24)	(0.03)	(0.08)	(0.04)	(0.18)	(0.14)	(0.26)	(0.04)	(0.05)	(0.07)
	A			B			C			D		
C	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.55	0.37	0.49	0.86	0.71	0.85	**0.91**	0.76	0.86	0.42	0.27	0.53
	(0.27)	(0.23)	(0.31)	(0.19)	(0.19)	(0.09)	(0.06)	(0.10)	(0.09)	(0.35)	(0.24)	(0.31)
REF + DA	0.65	0.47	0.38	0.89	0.77	0.75	0.89	**0.76**	0.86	0.63	0.43	0.21
	(0.29)	(0.27)	(0.32)	(0.09)	(0.08)	(0.21)	(0.09)	(0.12)	(0.08)	(0.31)	(0.25)	(0.29)
ASDNet	0.78	0.57	0.70	0.93	0.79	0.86	0.89	0.73	**0.86**	0.78	0.53	**0.73**
	(0.23)	(0.21)	(0.25)	(0.03)	(0.06)	(0.07)	(0.10)	(0.14)	(0.06)	(0.17)	(0.18)	(0.14)
Ours	**0.82**	**0.67**	**0.69**	**0.93**	**0.80**	**0.87**	0.90	0.75	0.85	0.86	0.65	0.71
	(0.23)	(0.20)	(0.30)	(0.03)	(0.05)	(0.05)	(0.09)	(0.12)	(0.06)	(0.05)	(0.08)	(0.19)
Ours + DA	0.75	0.60	0.61	0.91	0.78	0.82	0.87	0.69	0.80	**0.87**	**0.71**	0.72
	(0.22)	(0.19)	(0.28)	(0.03)	(0.04)	(0.07)	(0.09)	(0.13)	(0.08)	(0.05)	(0.08)	(0.09)
	A			B			C			D		
D	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.73	0.59	0.63	0.91	0.78	0.84	0.78	0.62	0.54	0.94	0.83	**0.92**
	(0.21)	(0.18)	(0.34)	(0.05)	(0.05)	(0.16)	(0.22)	(0.22)	(0.38)	(0.04)	(0.03)	(0.03)
REF + DA	**0.89**	**0.73**	0.78	0.91	0.77	**0.88**	**0.91**	0.73	**0.81**	0.94	**0.83**	0.90
	(0.07)	(0.07)	(0.12)	(0.04)	(0.06)	(0.08)	(0.05)	(0.11)	(0.07)	(0.03)	(0.02)	(0.04)
ASDNet	0.76	0.68	0.77	0.87	0.70	0.82	0.72	0.61	0.66	0.92	0.74	0.88
	(0.18)	(0.08)	(0.18)	(0.05)	(0.06)	(0.13)	(0.22)	(0.14)	(0.30)	(0.05)	(0.06)	(0.04)
Ours	0.86	0.69	0.67	0.90	0.78	0.85	0.86	**0.74**	0.68	0.93	0.81	0.90
	(0.11)	(0.10)	(0.29)	(0.04)	(0.04)	(0.10)	(0.08)	(0.09)	(0.30)	(0.03)	(0.03)	(0.03)
Ours + DA	0.88	0.70	**0.79**	**0.91**	**0.78**	0.85	0.83	0.69	0.64	**0.94**	0.82	0.90
	(0.10)	(0.12)	(0.16)	(0.03)	(0.04)	(0.07)	(0.19)	(0.14)	(0.25)	(0.03)	(0.03)	(0.04)

Table 4.2. Quantitative evaluations reporting mean Dice score and standard deviation (in parenthesis) for ES.

	A			B			C			D		
A	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.86	0.78	0.80	0.83	0.71	0.74	0.73	0.60	0.65	**0.83**	0.71	0.74
	(0.15)	(0.15)	(0.10)	(0.10)	(0.15)	(0.11)	(0.27)	(0.25)	(0.25)	(0.10)	(0.11)	(0.12)
REF + DA	**0.90**	**0.82**	**0.82**	**0.83**	0.73	**0.79**	0.74	0.65	**0.77**	0.81	0.75	**0.81**
	(0.08)	(0.09)	(0.10)	(0.08)	(0.17)	(0.11)	(0.19)	(0.20)	(0.09)	(0.12)	(0.08)	(0.10)
ASDNet	0.82	0.76	0.78	0.66	0.68	0.74	0.62	0.52	0.58	0.66	0.63	0.74
	(0.10)	(0.09)	(0.11)	(0.13)	(0.06)	(0.11)	(0.17)	(0.19)	(0.30)	(0.18)	(0.09)	(0.15)
Ours	0.87	0.81	0.78	0.81	**0.82**	0.74	**0.77**	**0.73**	0.64	0.79	**0.80**	0.79
	(0.08)	(0.06)	(0.09)	(0.10)	(0.05)	(0.11)	(0.12)	(0.09)	(0.25)	(0.10)	(0.04)	(0.08)
Ours + DA	0.86	0.80	0.76	0.72	0.73	0.63	0.72	0.63	0.55	0.79	0.74	0.74
	(0.09)	(0.06)	(0.11)	(0.13)	(0.05)	(0.12)	(0.11)	(0.13)	(0.22)	(0.10)	(0.05)	(0.11)
	A			B			C			D		
B	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.51	0.43	0.37	0.87	0.85	0.82	0.74	0.71	0.67	**0.86**	0.82	0.74
	(0.31)	(0.29)	(0.28)	(0.12)	(0.13)	(0.15)	(0.19)	(0.15)	(0.27)	(0.06)	(0.09)	(0.23)
REF + DA	0.45	0.42	0.27	0.87	0.85	**0.82**	**0.80**	**0.78**	**0.79**	0.84	0.78	0.68
	(0.36)	(0.32)	(0.28)	(0.08)	(0.10)	(0.13)	(0.13)	(0.08)	(0.11)	(0.12)	(0.18)	(0.23)
ASDNet	0.67	0.59	0.53	**0.88**	**0.87**	0.81	0.78	0.71	0.67	0.84	**0.82**	0.78
	(0.26)	(0.26)	(0.27)	(0.09)	(0.09)	(0.13)	(0.15)	(0.16)	(0.26)	(0.06)	(0.07)	(0.10)
Ours	0.68	0.52	0.54	0.77	0.82	0.70	0.68	0.67	0.51	0.76	0.77	0.66
	(0.25)	(0.24)	(0.23)	(0.11)	(0.09)	(0.09)	(0.27)	(0.26)	(0.21)	(0.13)	(0.10)	(0.13)
Ours + DA	**0.78**	**0.71**	**0.65**	0.84	0.85	0.78	0.72	0.70	0.64	0.81	0.81	**0.79**
	(0.23)	(0.22)	(0.23)	(0.09)	(0.05)	(0.06)	(0.27)	(0.24)	(0.25)	(0.07)	(0.05)	(0.07)
	A			B			C			D		
C	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.66	0.52	0.51	0.82	0.75	**0.73**	**0.86**	**0.81**	**0.79**	0.54	0.40	0.50
	(0.30)	(0.25)	(0.25)	(0.11)	(0.17)	(0.21)	(0.13)	(0.09)	(0.08)	(0.34)	(0.31)	(0.27)
REF + DA	0.71	0.62	0.42	0.77	0.73	0.66	0.81	0.77	0.73	0.58	0.50	0.36
	(0.27)	(0.22)	(0.27)	(0.18)	(0.19)	(0.28)	(0.12)	(0.12)	(0.13)	(0.31)	(0.29)	(0.34)
ASDNet	0.78	0.66	0.61	**0.85**	0.81	0.72	0.83	0.79	0.71	0.76	0.64	0.71
	(0.27)	(0.23)	(0.26)	(0.09)	(0.13)	(0.12)	(0.16)	(0.12)	(0.07)	(0.15)	(0.19)	(0.20)
Ours	**0.79**	**0.68**	**0.64**	0.84	0.81	0.72	0.82	0.78	0.71	0.80	0.70	**0.72**
	(0.23)	(0.22)	(0.27)	(0.08)	(0.12)	(0.12)	(0.15)	(0.13)	(0.08)	(0.12)	(0.15)	(0.11)
Ours + DA	0.72	0.65	0.57	0.77	**0.81**	0.67	0.72	0.72	0.63	**0.82**	**0.76**	0.68
	(0.25)	(0.22)	(0.29)	(0.10)	(0.10)	(0.12)	(0.15)	(0.13)	(0.11)	(0.05)	(0.09)	(0.14)
	A			B			C			D		
D	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
REF	0.72	0.62	0.57	0.81	0.76	0.71	0.73	0.65	0.53	0.94	0.88	**0.88**
	(0.25)	(0.23)	(0.32)	(0.12)	(0.19)	(0.26)	(0.31)	(0.29)	(0.36)	(0.02)	(0.03)	(0.08)
REF + DA	**0.83**	**0.74**	**0.73**	**0.83**	**0.82**	**0.77**	**0.81**	**0.74**	**0.70**	**0.94**	**0.89**	0.86
	(0.16)	(0.13)	(0.16)	(0.09)	(0.06)	(0.14)	(0.12)	(0.17)	(0.17)	(0.03)	(0.02)	(0.07)
ASDNet	0.81	0.71	0.63	0.73	0.72	0.68	0.68	0.62	0.62	0.86	0.79	0.85
	(0.10)	(0.08)	(0.28)	(0.11)	(0.10)	(0.23)	(0.27)	(0.27)	(0.28)	(0.10)	(0.06)	(0.05)
Ours	0.80	0.71	0.59	0.73	0.80	0.70	0.70	0.71	0.61	0.89	0.85	0.87
	(0.21)	(0.14)	(0.30)	(0.11)	(0.04)	(0.17)	(0.19)	(0.19)	(0.30)	(0.04)	(0.04)	(0.04)
Ours + DA	0.80	0.69	0.68	0.78	0.81	0.72	0.69	0.64	0.52	0.91	0.87	0.86
	(0.20)	(0.19)	(0.21)	(0.09)	(0.06)	(0.10)	(0.28)	(0.27)	(0.28)	(0.03)	(0.03)	(0.06)

Chapter 5

Conclusions

This thesis yielded two algorithms as an attempt to solve respectively the problem of generalization and the problem of automatic quality assessment in the field of CMR segmentation. Both the proposed methods revolve around the development of a convolutional autoencoder, which provides segmentation masks clean of peculiar defects with reference to any training dataset provided with reliable ground truth. We called the output of the autoencoder pseudo Ground Truth (pGT), and we used it to derive quality measures in Chapter 3 and to develop a semi-supervised framework in Chapter 4.

We provided two different types of quality measures, a global score, and a pixel-wise map. The former acts as a monitor to flag irregularities; the latter locates the cause of a low-quality verdict inside the mask itself. Combined, the two measures allow surveilling the performance of cardiac image segmentation models in the absence of ground truth. Compared with previous approaches, our method is more robust, scalable, and considerably fast. This permits a quick quality assignment, for example, in a clinical setting.

When plugged into a semi-supervised framework, our autoencoder provides a reference to calculate a dissimilarity loss function on unlabeled data. These data do not need additional labeling effort but can provide a large representation of the cardiac anatomy, improving cross-domain generalisation. Experiments led to incorporate a QC module into the learning pipeline of a segmentation model. Working together, these two modules achieved state-of-the-art performances in cardiac image segmentation. Future experiments should investigate the possibility of using the QC module to select the data from the target domain with a real need to refine the model. In addition, our work focuses on the problem of domain shift within a single modality, which can be considered as a limitation. Future research could be directed towards the extension of the method to multiple modalities, which existing QC methods cannot cope with.

Bibliography

[1] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. M. Jodoin, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?", IEEE Transactions on Medical Imaging, vol. 37, no. 11, 2018, pp. 2514–2525, DOI 10.1109/TMI.2018.2837502

[2] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, "Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age", PLOS Medicine, vol. 12, 03 2015, pp. 1–10, DOI 10.1371/journal.pmed.1001779

[3] A. H. Allam, R. C. Thompson, L. S. Wann, M. I. Miyamoto, A. el-Halim Nur el Din, G. A. el Maksoud, M. A.-T. Soliman, I. Badr, H. A. el Rahman Amer, M. L. Sutherland, J. D. Sutherland, and G. S. Thomas, "Atherosclerosis in ancient egyptian mummies", JACC: Cardiovascular Imaging, vol. 4, no. 4, 2011, pp. 315–327, DOI 10.1016/j.jcmg.2011.02.002

[4] W. C. Lam and D. J. Pennell, "Imaging of the heart: historical perspective and recent advances", Postgraduate Medical Journal, vol. 92, no. 1084, 2016, pp. 99–104, DOI 10.1136/postgradmedj-2015-133831

[5] C. Chow and T. Kaneko, "Automatic boundary detection of the left ventricle from cineangiograms", Computers and Biomedical Research, vol. 5, no. 4, 1972, pp. 388 – 410, DOI https://doi.org/10.1016/0010-4809(72)90070-5

[6] Y. Nakagawa and A. Rosenfeld, "Some experiments on variable thresholding", Pattern Recognition, vol. 11, no. 3, 1979, pp. 191 – 204, DOI https://doi.org/10.1016/0031-3203(79)90006-2

[7] S. Yanowitz and A. Bruckstein, "A new method for image segmentation", Computer Vision, Graphics, and Image Processing, vol. 46, no. 1, 1989, pp. 82 – 95, DOI https://doi.org/10.1016/S0734-189X(89)80017-9

[8] K. V. Mardia and T. J. Hainsworth, "A spatial thresholding method for image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 10, no. 6, 1988, pp. 919–927, DOI 10.1109/34.9113

[9] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method", IEEE Transactions on Systems, Man, and Cybernetics, vol. 8, no. 8,

1978, pp. 630–632, DOI 10.1109/TSMC.1978.4310039

[10] P. K. Sahoo, S. Soltani, A. K. C. Wong, and Y. C. Chen, "A survey of thresholding techniques", Computer Vision, Graphics, and Image Processing, vol. 41, no. 2, 1988, pp. 233 – 260, DOI https://doi.org/10.1016/0734-189X(88)90022-9

[11] N. R. Pal and S. K. Pal, "A review on image segmentation techniques", Pattern Recognition, vol. 26, no. 9, 1993, pp. 1277 – 1294, DOI https://doi.org/10.1016/0031-3203(93)90135-J

[12] A. Rosenfeld, R. A. Hummel, and S. W. Zucker, "Scene labeling by relaxation operations", IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 6, 1976, pp. 420–433, DOI 10.1109/TSMC.1976.4309519

[13] S. Peleg, "A new probabilistic relaxation scheme", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. PAMI-2, no. 4, 1980, pp. 362–369, DOI 10.1109/TPAMI.1980.4767035

[14] L. S. Davis, "A survey of edge detection techniques", Computer Graphics and Image Processing, vol. 4, no. 3, 1975, pp. 248 – 270, DOI https://doi.org/10.1016/0146-664X(75)90012-X

[15] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active shape models-their training and application", Computer Vision and Image Understanding, vol. 61, no. 1, 1995, pp. 38 – 59, DOI https://doi.org/10.1006/cviu.1995.1004

[16] 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, 2004, pp. 255 – 265, DOI https://doi.org/10.1016/j.media.2004.06.005. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003

[17] A. Andreopoulos and J. K. Tsotsos, "Efficient and generalizable statistical models of shape and appearance for analysis of cardiac mri", Medical Image Analysis, vol. 12, no. 3, 2008, pp. 335 – 357, DOI https://doi.org/10.1016/j.media.2007.12.003

[18] 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, 2013, pp. 966 – 989, DOI https://doi.org/10.1016/j.cviu.2012.11.017

[19] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer, "Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains", NeuroImage, vol. 21, no. 4, 2004, pp. 1428 – 1442, DOI https://doi.org/10.1016/j.neuroimage.2003.11.010

[20] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain mri segmentation combining label propagation and decision fusion", NeuroImage, vol. 33, no. 1, 2006, pp. 115 – 126, DOI https://doi.org/10.1016/j.neuroimage.2006.05.061

[21] 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, 2009, pp. 1000–1010, DOI 10.1109/TMI.2008.2011480

[22] 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", Functional Imaging and Modeling of the Heart, 2013, pp. 174–181, DOI 10.1007/978-3-642-38899-6_21

[23] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models", Computer Vision — ECCV'98 (H. Burkhardt and B. Neumann, eds.), Berlin, Heidelberg, 1998, pp. 484–498

[24] J. Hansegard, S. Urheim, K. Lunde, and S. I. Rabben, "Constrained active appearance models for segmentation of triplane echocardiograms", IEEE Transactions on Medical Imaging, vol. 26, no. 10, 2007, pp. 1391–1400, DOI 10.1109/TMI.2007.900692

[25] H. Zhang, A. Wahle, R. K. Johnson, T. D. Scholz, and M. Sonka, "4-d cardiac mr image analysis: Left and right ventricular morphology and function", IEEE Transactions on Medical Imaging, vol. 29, no. 2, 2010, pp. 350–364, DOI 10.1109/TMI.2009.2030799

[26] I. Dydenko, F. Jamal, O. Bernard, J. D'hooge, 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, 2006, pp. 162 – 177, DOI https://doi.org/10.1016/j.media.2005.06.004

[27] K. Punithakumar, I. B. Ayed, A. Islam, I. G. Ross, and S. Li, "Tracking endocardial motion via multiple model filtering", IEEE Transactions on Biomedical Engineering, vol. 57, no. 8, 2010, pp. 2001–2010, DOI 10.1109/TBME.2010.2048752

[28] F. Isensee, P. F. 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", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 120–129, DOI 10.1007/978-3-319-75541-0_13

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), Cham, 2015, pp. 234–241, DOI 10.1007/978-3-319-24574-4_28

[30] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), Cham, 2016, pp. 424–432, DOI 10.1007/978-3-319-46723-8_49

[31] 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", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 111–119, DOI 10.1007/978-3-319-75541-0_12

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440, DOI 10.1109/CVPR.2015.7298965

[33] J. Patravali, S. Jain, and S. Chilamkurthy, "2d-3d fully convolutional neural networks for cardiac mr segmentation", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant,

P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 130–139, DOI 10.1007/978-3-319-75541-0_14

[34] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations", Computer Vision, Graphics, and Image Processing, vol. 39, no. 3, 1987, pp. 355 – 368, DOI https://doi.org/10.1016/S0734-189X(87)80186-X

[35] X. Yang, C. Bian, L. Yu, D. Ni, and P.-A. Heng, "Class-balanced deep neural network for automatic ventricular structure segmentation", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 152–160, DOI 10.1007/978-3-319-75541-0_16

[36] Y. Jang, Y. Hong, S. Ha, S. Kim, and H.-J. Chang, "Automatic segmentation of lv and rv in cardiac mri", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 161–169, DOI 10.1007/978-3-319-75541-0_17

[37] R. Mehta and J. Sivaswamy, "M-net: A convolutional neural network for deep brain structure segmentation", 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017, pp. 437–440, DOI 10.1109/ISBI.2017.7950555

[38] M. Khened, V. Alex, and G. Krishnamurthi, "Densely connected fully convolutional network for short-axis cardiac cine mr image segmentation and heart diagnosis using random forest", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 140–151, DOI 10.1007/978-3-319-75541-0_15

[39] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation", 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1175–1183, DOI 10.1109/CVPRW.2017.156

[40] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions", 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9, DOI 10.1109/CVPR.2015.7298594

[41] M.-M. Rohé, M. Sermesant, and X. Pennec, "Automatic multi-atlas segmentation of myocardium with svf-net", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 170–177, DOI 10.1007/978-3-319-75541-0_18

[42] M.-M. Rohé, M. Datar, T. Heimann, M. Sermesant, and X. Pennec, "Svf-net: Learning deformable image registration using shape matching", Medical Image Computing and Computer Assisted Intervention MICCAI 2017 (M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, eds.), Cham, 2017, pp. 266–274, DOI 10.1007/978-3-319-66182-7_31

[43] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P.-M. Jodoin, "Gridnet with automatic shape prior registration for automatic mri cardiac segmentation", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 73–81, DOI 10.1007/978-3-319-75541-0_8

[44] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Automatic segmentation and disease classification using cardiac cine mr images", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 101–110, DOI 10.1007/978-3-319-75541-0_11

[45] E. Grinias and G. Tziritas, "Fast fully-automatic cardiac segmentation in mri using mrf model optimization, substructures tracking and b-spline smoothing", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 91–100, DOI 10.1007/978-3-319-75541-0_10

[46] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, F. Zemrak, K. Fung, J. M. Paiva, V. Carapella, Y. J. Kim, H. Suzuki, B. Kainz, P. M. Matthews, S. E. Petersen, S. K. Piechnik, S. Neubauer, B. Glocker, and D. Rueckert, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks", Journal of Cardiovascular Magnetic Resonance, vol. 20, no. 1, 2018, p. 65, DOI 10.1186/s12968-018-0471-x

[47] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison, "Hidden technical debt in machine learning systems", Advances in Neural Information Processing Systems (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), 2015, pp. 2503–2511, DOI 10.5555/2969442.2969519

[48] G.Tarroni, W.Bai, and O.Oktay, "Large-scale Quality Control of Cardiac Imaging in Population Studies: Application to UK Biobank.", Sci Rep, vol. 10, February 2020, DOI 10.1038/s41598-020-58212-2

[49] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, "Evaluating segmentation error without ground truth", Medical Image Computing and Computer-Assisted Intervention – MICCAI 2012 (N. Ayache, H. Delingette, P. Golland, and K. Mori, eds.), Berlin, Heidelberg, 2012, pp. 528–536, DOI 10.1007/978-3-642-33415-3_65

[50] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, B. Kainz, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, D. Rueckert, e. A. F. Glocker, Ben", J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, "Real-Time Prediction of Segmentation Quality.", Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, pp. 578–585, Springer International Publishing, 2010, DOI 10.1007/978-3-030-00937-3_66

[51] E. Puyol-Antón, B. Ruijsink, C. Baumgartner, P. Masci, M. Sinclair,

E. Konukoglu, R. Razavi, and A. King, "Automated quantification of myocardial tissue characteristics from native T_1 mapping using neural networks with uncertainty-based quality-control", Journal of Cardiovascular Magnetic Resonance, vol. 22, August 2020, DOI 10.1186/s12968-020-00650-y

[52] R. Robinson, V. V. Valindria, W. Bai, O. Oktay, B. Kainz, H. Suzuki, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, P. M. Matthews, D. Rueckert, and B. Glocker, "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study", Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance, vol. 21, March 2019, p. 18, DOI 10.1186/s12968-019-0523-x

[53] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, "Reverse classification accuracy: Predicting segmentation performance in the absence of ground truth", IEEE Transactions on Medical Imaging, vol. 36, no. 8, 2017, pp. 1597–1606, DOI 10.1109/TMI.2017.2665165

[54] M. A. Zuluaga, N. Burgos, A. F. Mendelson, A. M. Taylor, and S. Ourselin, "Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries", Medical Image Analysis, vol. 26, no. 1, 2015, pp. 185–194, DOI 10.1016/j.media.2015.09.001

[55] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "MVTec AD - A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection", 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019, pp. 9584–9592, DOI 10.1109/CVPR.2019.00982

[56] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning", MIT Press, 2016. http://www.deeplearningbook.org

[57] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", Advances in Neural Information Processing Systems 27 (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2672–2680, Curran Associates, Inc., 2014. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf

[58] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series", Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining, New York, NY, USA, 2020, p. 3395–3404, DOI 10.1145/3394486.3403392

[59] P. Bergmann., S. Löwe., M. Fauser., D. Sattlegger., and C. Steger., "Improving unsupervised defect segmentation by applying structural similarity to autoencoders", Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP, 2019, pp. 372–380, DOI 10.5220/0007364503720380

[60] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations", Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, A. Bradley, H. Greenspan,

J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, and Z. Lu, eds.), Cham, 2017, pp. 240–248, DOI 10.1007/978-3-319-67558-9_28

[61] C. Zotti, Z. Luo, O. Humbert, A. Lalande, and P.-M. Jodoin, "Gridnet with automatic shape prior registration for automatic mri cardiac segmentation", Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges (M. Pop, M. Sermesant, P.-M. Jodoin, A. Lalande, X. Zhuang, G. Yang, A. Young, and O. Bernard, eds.), Cham, 2018, pp. 73–81, DOI 10.1007/978-3-319-75541-0_8

[62] V. M. Campello, J. F. R. Palomares, A. Guala, M. Marakas, M. Friedrich, and K. Lekadir, "Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge", March 2020, DOI 10.5281/zenodo.3886268

[63] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy", 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8049–8057, DOI 10.1109/ICCV.2019.00814

[64] A. Torralba and A. A. Efros, "Unbiased look at dataset bias", CVPR 2011, 2011, pp. 1521–1528, DOI 10.1109/CVPR.2011.5995347

[65] C. Chen, W. Bai, R. H. Davies, A. N. Bhuva, C. H. Manisty, J. B. Augusto, J. C. Moon, N. Aung, A. M. Lee, M. M. Sanghvi, K. Fung, J. M. Paiva, S. E. Petersen, E. Lukaschuk, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Improving the generalizability of convolutional neural network-based segmentation on cmr images", Frontiers in Cardiovascular Medicine, vol. 7, 2020, p. 105, DOI 10.3389/fcvm.2020.00105

[66] J. Chen, H. Zhang, Y. Zhang, S. Zhao, R. Mohiaddin, T. Wong, D. Firmin, G. Yang, and J. Keegan, "Discriminative consistent domain generation for semi-supervised learning", Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), Cham, 2019, pp. 595–604, DOI 10.1007/978-3-030-32245-8_66

[67] F. Guo, M. Ng, M. Goubran, S. E. Petersen, S. K. Piechnik, S. Neubauer, and G. Wright, "Improving cardiac mri convolutional neural network segmentation on small training datasets and dataset shift: A continuous kernel cut approach", Medical Image Analysis, vol. 61, 2020, p. 101636, DOI https://doi.org/10.1016/j.media.2020.101636

[68] M. Khened, V. A. Kollerathu, and G. Krishnamurthi, "Fully convolutional multi-scale residual densenets for cardiac segmentation and automated cardiac diagnosis using ensemble of classifiers", Medical Image Analysis, vol. 51, 2019, pp. 21 – 45, DOI https://doi.org/10.1016/j.media.2018.10.004

[69] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert, "Data efficient unsupervised domain adaptation for cross-modality image segmentation", Medical Image Computing and Computer Assisted Intervention – MICCAI 2019 (D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, eds.), Cham, 2019, pp. 669–677, DOI 10.1007/978-3-030-32245-8_74

[70] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation", IEEE Transactions on Medical Imaging, vol. 39, no. 7, 2020, pp. 2494–2505, DOI 10.1109/TMI.2020.2972701

[71] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation", Medical Image Computing and Computer Assisted Intervention – MICCAI 2018 (A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, eds.), Cham, 2018, pp. 370–378, DOI 10.1007/978-3-030-00937-3_43

[72] W. Hung, Y. Tsai, Y. Liou, Y. Lin, and M. Yang, "Adversarial learning for semi-supervised semantic segmentation", CoRR, vol. abs/1802.07934, 2018

[73] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks", 5th International Conference on Learning Representations, ICLR 2017, OpenReview.net, Toulon, France, 2017

[74] C. Chen, C. Qin, H. Qiu, G. Tarroni, J. Duan, W. Bai, and D. Rueckert, "Deep learning for cardiac image segmentation: A review", Frontiers in Cardiovascular Medicine, vol. 7, 2020, p. 25, DOI 10.3389/fcvm.2020.00025

[75] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of neural networks using dropconnect", International conference on machine learning, 2013, pp. 1058–1066

[76] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1026–1034, DOI 10.1109/ICCV.2015.123