THE UNIVERSITY
of EDINBURGH

Multimodal and Disentangled Representation Learning for Medical Image Analysis

Agisilaos Chartsias

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
2020

Abstract

Automated medical image analysis is a growing research field with various applications in modern healthcare. Furthermore, a multitude of imaging techniques (or modalities) have been developed, such as Magnetic Resonance (MR) and Computed Tomography (CT), to attenuate different organ characteristics. Research on image analysis is predominately driven by deep learning methods due to their demonstrated performance. In this thesis, we argue that their success and generalisation relies on learning good latent representations. We propose methods for learning spatial representations that are suitable for medical image data, and can combine information coming from different modalities. Specifically, we aim to improve cardiac MR segmentation, a challenging task due to varied images and limited expert annotations, by considering complementary information present in (potentially unaligned) images of other modalities.

In order to evaluate the benefit of multimodal learning, we initially consider a synthesis task on spatially aligned multimodal brain MR images. We propose a deep network of multiple encoders and decoders, which we demonstrate outperforms existing approaches. The encoders (one per input modality) map the multimodal images into modality invariant spatial feature maps. Common and unique information is combined into a fused representation, that is robust to missing modalities, and can be decoded into synthetic images of the target modalities. Different experimental settings demonstrate the benefit of multimodal over unimodal synthesis, although input and output image pairs are required for training. The need for paired images can be overcome with the cycle consistency principle, which we use in conjunction with adversarial training to transform images from one modality (e.g. MR) to images in another (e.g. CT). This is useful especially in cardiac datasets, where different spatial and temporal resolutions make image pairing difficult, if not impossible.

Segmentation can also be considered as a form of image synthesis, if one modality consists of semantic maps. We consider the task of extracting segmentation masks for cardiac MR images, and aim to overcome the challenge of limited annotations, by taking into account unannanotated images which are commonly ignored. We achieve this by defining suitable latent spaces, which represent the underlying anatomies (spatial latent variable), as well as the imaging characteristics (non-spatial latent variable). Anatomical information is required for tasks such as segmentation and regression, whereas imaging information can capture variability in intensity

characteristics for example due to different scanners. We propose two models that disentangle cardiac images at different levels: the first extracts the myocardium from the surrounding information, whereas the second fully separates the anatomical from the imaging characteristics. Experimental analysis confirms the utility of disentangled representations in semi-supervised segmentation, and in regression of cardiac indices, while maintaining robustness to intensity variations such as the ones induced by different modalities.

Finally, our prior research is aggregated into one framework that encodes multimodal images into disentangled anatomical and imaging factors. Several challenges of multimodal cardiac imaging, such as input misalignments and the lack of expert annotations, are successfully handled in the shared anatomy space. Furthermore, we demonstrate that this approach can be used to combine complementary anatomical information for the purpose of multimodal segmentation. This can be achieved even when no annotations are provided for one of the modalities.

This thesis creates new avenues for further research in the area of multimodal and disentangled learning with spatial representations, which we believe are key to more generalised deep learning solutions in healthcare.

Lay Summary

Medical imaging is widely used for the diagnosis and treatment of different pathological conditions. Many techniques can image internal organs, for example Magnetic Resonance (MR) uses the tissue magnetic properties, and Computed Tomography (CT) uses X-Rays. Each technique (also known as modality) has its own characteristics, produces grayscale images of different brightness (intensity) and enhances the contrast of organs and pathology differently. For instance within cardiac MR, cine-MR creates a "movie" of the moving heart and is used to assess the cardiac function, and Latent Gadolinium Enhancement (LGE) uses an injected paramagnetic substance that enhances the contrast of infarcted regions of the heart muscle (myocardium), i.e. regions with reduced blood flow that cause heart attack. Typically, the analysis of such images is a manual process that is time consuming and requires expertise. This entails delineating the position of the myocardium within the image (annotation) by experts, followed by a quantitative analysis of the cardiac function. There is therefore a need for automated methods that can alleviate the requirement (as well as reduce the cost) for myocardium annotations.

In recent years, many methods for automating image analysis tasks have been proposed. These primarily belong to a class of so-called "deep learning" models, which "learn" to perform a particular task by using pairs of input and output examples. In this thesis, we aim to develop deep learning models to extract myocardial delineations from input images, a task termed as segmentation. The development of such models is split in two stages: learning and inference. During learning, the models are "trained" to perform the task of segmentation using examples of images with their corresponding annotations. During inference, the models predict segmentations when given new unseen images. We further aim to combine information present in images of different modalities (multimodal) in order to improve the accuracy of predicted segmentations. This is challenging, especially in cardiac images, because of differences in intensity characteristics between the modalities, and variation in anatomy (the heart is a moving organ).

In order to evaluate the benefits of multimodal learning, we initially consider multimodal brain MR images, because they are always aligned (unlike the heart, the brain does not move). We propose a method that takes as input a brain image in some input modalities and produces the same image in an output modality. This is known as synthesis. We achieve this task with a deep

learning model that firstly transforms the intensities of the multimodal inputs to be similar, so that images can be directly compared and combined, and secondly transforms the intensities of the combined images to correspond to the output modality. This synthesis model is then extended to also work with multimodal inputs that are not aligned, and therefore it can be applied to cardiac images. This extended model does not require having the same image in two modalities and can learn with any multimodal data, for instance with data of MR and CT images of different patients.

We further enhance the performance of deep learning methods by devising an intuitive disentanglement (or decomposition) of medical images in two factors. The first corresponds to the underlying anatomy that is common across all modalities, for example the heart, and the second to the modality characteristics that are common across all anatomies, for example the range of grayscale intensity values in MR imaging. Therefore, any image analysis task, such as segmentation, can be performed by only using the anatomy factors. In addition, such a disentanglement reduces the requirement for having many expert annotations, a critical limitation in medical imaging. They further enable multimodal processing by combining anatomies produced by images in different modalities. Our final framework is able to combine multimodal cardiac images by first disentangling them in their corresponding factors. Extensive experiments demonstrate accurate segmentation results when limited amount or no annotated data are provided for one of the modalities.

This thesis creates new avenues for further research in the area of multimodal processing and image decompositions, which we believe are key to more generalised deep learning solutions in healthcare.

Acknowledgements

I would like to express my special appreciation to my advisor Prof. Sotirios Tsaftaris. Thank you Sotos for your valued guidance throughout these years and for sharing your passion for research. Your hard working spirit and insights to the field have been inspiring. I will always remember our fruitful conversations and the good times we had in our various conference trips. A big thank you to Dr. Thomas Joyce. I was very lucky to collaborate with such a smart and kind person. I also thank Dr. Javier Escudero and Prof. Mike Davies for their useful feedback throughout the course of my PhD. I thank Dr. Giorgos Papanastasiou, Dr. Chengjia Wang for our discussions and along with the collaborators in the Royal Infirmary for providing experimental data. I thank Dr. Antonis Giannopoulos for giving me access to GPUs for my experiments. Both resources were invaluable to this thesis. I thank all my colleagues, Andrei, Gabriele, Greg, Haochuan, Heyi, Ilkay, Spyro, Tian, Valerio, and Xiao for creating a great atmosphere in the office. Working alongside you all has been a joy. I thank all my friends for all enjoyable moments outside PhD life. I cannot thank enough my mother, father and sister for their love and continuous support all these years. You have made me the person I am today. Finally, my special thanks go to my other half, my beloved Thalia, who has been by my side since the beginning, throughout all highs and lows. Without you none of this would have been possible.

List of Publications

Thesis publications:

- Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2019. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE Transactions on Medical Imaging (under review).
  Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Papanastasiou, G.: data curation, resources, validation, writing - review & editing; Wang, C.: conceptualisation, validation; Semple, S.: resources; Newby, D.E.: resources; writing - review & editing; Dharmakumar, R.: funding acquisition, resources; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
  Article is presented in Chapter 7.

- Chartsias, A., Papanastasiou, G., Wang, C., Stirrat, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2019. Multimodal Cardiac Segmentation Using Disentangled Representation Learning. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 128-137). Springer, Cham.
  Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Papanastasiou, G.: data curation, resources, validation, writing - review & editing; Wang, C.: conceptualisation, validation; Stirrat, C.: resources; Semple, S.: resources; Newby, D.E.: resources; Dharmakumar, R.: funding acquisition, resources; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
  Article presented in Chapter 7.

- Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R. and Tsaftaris, S.A., 2019. Disentangled representation learning in cardiac image analysis. Medical Image Analysis, 58, p.101535.

Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Joyce, T.: conceptualisation, validation, visualisation, writing - review & editing; Papanastasiou, G.: data curation, resources, validation, writing - review & editing; Semple, S.: resources; Williams, M.: resources; Newby, D.E.: resources; Dharmakumar, R.: funding acquisition, resources; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
Article presented in Chapter 6.

- Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2018. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 490-498). Springer, Cham.
  Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Joyce, T.: conceptualisation, validation, visualisation, writing - review & editing; Papanastasiou, G.: data curation, resources, validation; Semple, S.: resources; Williams, M.: resources; Newby, D.E.: resources; Dharmakumar, R.: funding acquisition; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
  Article presented in Chapter 6.

- Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017, September. Adversarial image synthesis for unpaired multi-modal cardiac data. In International workshop on simulation and synthesis in medical imaging (pp. 3-13). Springer, Cham.
  Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Joyce, T.: conceptualisation, data curation, formal analysis, methodology, software, validation, visualisation, writing - original draft, writing - review & editing; Dharmakumar, R.: funding acquisition; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
  Article presented in Chapter 5.

- Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A., 2018. Multimodal MR synthesis via modality-invariant latent representation. IEEE Transactions on Medical Imaging, 37(3), pp.803-814.

  Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Joyce, T.: conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualisation, writing - original draft, writing - review & editing; Giuffrida, M.V.: investigation, software, writing - review & editing; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
  Article presented in Chapter 4.

- Joyce, T., Chartsias, A., Tsaftaris, S.A., 2017. Robust multi-modal MR image synthesis. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 347-355). Springer, Cham.

  Author contributions. Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, software, validation, visualisation, writing - original draft, writing - review & editing; Joyce, T.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing; Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.
  Article presented in Chapter 4.

Contents

List of figures

List of tables

Acronyms and Abbreviations

AO	ascending AOrta
BOLD	Blood Oxygen Level Dependent
CE	Cross Entropy
CMR	Cardiac Magnetic Resonance
CT	Computed Tomography
DAFNet	Disentangle Align and Fuse Network
DWI	Diffusion Weighted Image
ECG	Electrocardiogram
ED	End Diastole
ELBO	Evidence Lower Bound
ES	End Systole
FiLM	Feature-wise Linear Modulation
FLAIR	Fluid Attenuated Inversion Recovery
HeMIS	Hetero-Modal Image Segmentation
ISLES	Ischemic Stroke Lesion Segmentation
LA	Left Atrium
LGE	Late Gadolinium Enhancement
LV	Left Ventricle
LVV	Left Ventricular Volume
MAE	Mean Absolute Error
LSDN	Location Sensitive Deep Network
MMWHS	Multimodal Whole Heart Segmentation
MP	Modality Propagation
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MUNIT	Multimodal UNsupervised Image-to-image Translation
MYO	Myocardium
PD	Proton Density
RA	Right Atrium

REPLICA	Regression Ensembles with Patch Learning for Image Contrast Agreement
RV	Right Ventricle
GAN	Generative Adversarial Network
MR	Magnetic Resonance
PSNR	Peak Signal to Noise Ratio
ReLU	Rectified Linear Unit
RF	Radio Frequency
SDNet	Spatial Disentanglement Network
SMDNet	Spatial Myocardial Disentanglement Network
SPADE	Spatially-Adaptive (De)Normalization
SSIM	Structural Similarity Index
STN	Spatial Transformer Network
VAE	Variational Autoencoder

Nomenclature

N	number of image domains or modalities
H	image height
W	image width
C	number of image channels
V	number of segmentation mask channels
$c \in [1, C]$	channel index
n_z	number of dimensions of a vector z
$X \subset \mathbb{R}^{H \times W}$	set of images
$Y \subset \mathbb{R}^{H \times W}$	set of synthetic images
$M \coloneqq \{0, 1\}^{H \times W \times V}$	set of segmentation masks
$R \subset \mathbb{R}^{H \times W \times C}$	set of image representations
$S \coloneqq \{0, 1\}^{H \times W \times C}$	set of anatomical representations
$Z \coloneqq \mathbb{R}^{n_z}$	set of vector representations
x	image sample $x \in X$
y	image sample $y \in Y$
m	segmentation mask sample $m \in M$
r	spatial latent variable for an image representation $r \in R$
s	spatial latent variable for an anatomical representation $s \in S$
z	vectorised latent variable $z \in Z$
f	image encoder
g	image decoder
h	segmentation network
D	discriminator network
G	generator network
μ	mean
σ^2	variance

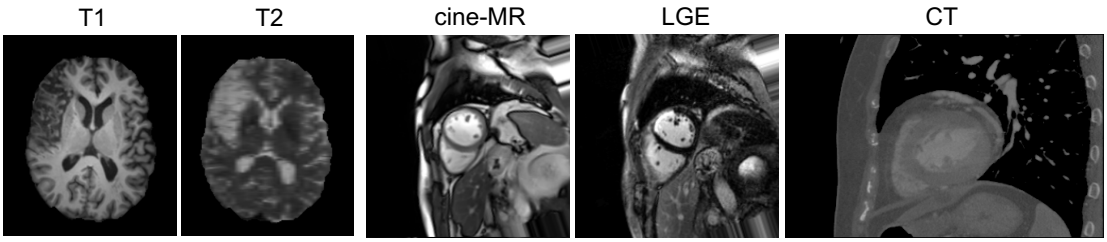
Chapter 1
Introduction

In deep learning, and medical image analysis in particular, learning good representations is key for developing solutions and offers many benefits. These include the ability to generalise to unseen data and related tasks, and also to learn smooth manifolds of the explanatory data factors [17]. We argue that learning *spatial* representations, i.e. tensors (feature maps) obtained from convolutional neural networks that are spatially co-variant with the input images, is crucial for developing automatic medical image analysis methods due to specific intricacies posed by medical data. In particular, we describe new methods for learning *multimodal* and *disentangled* spatial representations and demonstrate their utility in several medical applications.

1.1 Medical Motivation

In medical image analysis, learning multimodal and disentangled representations is an intuitive direction due to the nature of the medical imaging data and the challenges they present. Medical images are naturally multimodal, with modalities referring to either the different techniques, such as Magnetic Resonance (MR) and Computed Tomography (CT), or to different sequences within MR (multi-parametric), in which different settings or the use of contrast agents can accentuate T1 and T2 content in the imaged tissue. Example multimodal brain and cardiac images can be seen in Figure 1.1. Throughout this thesis, the term multimodal is used for both cases, although always clearly defined. Interest in multimodal images is high due to the complementary information that they encode for the underlying organs. For instance, multi-parametric MR is used in the brain for the detection of cancerous tumours [18], and in the heart for the assessment of cardiovascular status [19].

Multimodal image analysis is possible by learning spatial correlations across the modalities. However, although multimodal brain MR images are spatially aligned (see Figure 1.1a), this does not hold true for all organs. For instance, multimodal cardiac images often differ in spatial resolution with non-isotropic volumes that have different spacing among the slices. Furthermore, since the heart is a moving organ, multimodal images additionally differ in temporal



(a) multi-parametric T1 and T2 brain MR images

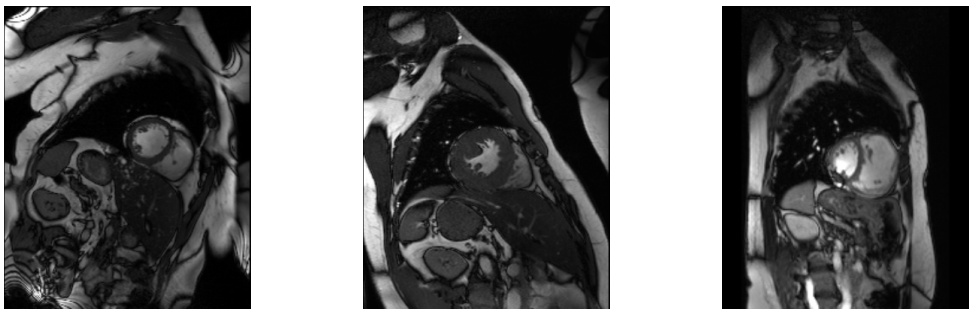
(b) multi-parametric MR cardiac images in cine-MR and LGE, and a cardiac CT image

Figure 1.1: Examples of multimodal brain and cardiac images. Brain images are from the same subject and are taken from Ischemic Stroke Lesion Segmentation (ISLES) dataset [1]. Cine-MR and LGE cardiac images are from the same subject, and are acquired as part of the study in [2]. CT image comes from a different subject and is taken from Multimodal Whole Heart Segmentation (MMWHS) dataset [3, 4]. Observe that multimodal brains are spatially aligned. Also, there is no pixel correspondence between the cardiac cine-MR and LGE images, although they both correspond to the same volume slice and diastolic frame.

resolution, making precise temporal alignment across sequences difficult (see Figure 1.1b). Therefore, multimodal alignment is required, prior to capturing the desired spatial correlations with information fusion techniques.

Cardiac MR image analysis, a primal focus of this thesis, presents further challenges. The heart shape exhibits great variability especially among patients with different pathological conditions, see for instance the images of Figure 1.2. Also, the intensities between adjacent sub-structures or tissues can be similar, for example between the myocardium and the papillary muscles, or between the heart and liver. Additional difficulties include image artefacts or further intensity inconsistencies across patients, pathologies, scanning sites and domain shift within modalities between devices, for example presented in qualitative MRI.

Specifically here we are interested in the task of myocardial segmentation. This has a great diagnostic value, because of the functional indices that can be calculated such as the ejection fraction, and the myocardial mass [20]. This segmentation task needs image annotations, which is a laborious and challenging task, also requiring medical expertise. The lack of annotations often results in small datasets, in which the proportion of unlabelled images is far higher than that of the labelled ones. This motivates research on semi-supervised approaches to achieve robust models, by taking into account unlabelled images. Approaches based on disentangled



(a) normal heart **(b)** hypertrophic cardiomyopathy **(c)** abnormal right ventricle

Figure 1.2: Three examples of cine-MR demonstrating differences in the cardiac anatomy due to pathological conditions. Images are taken from Automatic Cardiac Diagnosis Challenge (ACDC) dataset [5].

representations are suitable due to their inherent property of training with no supervision [21].

Based on the above considerations, we argue that multimodal and disentangled learning are valuable, both for leveraging information present in other modalities, and for utilising unlabelled images. Furthermore, segmentation is a task that is spatially equivariant with the input, meaning that spatial transformations to the images should also propagate to their corresponding masks. This motivates representing multimodal and disentangled latent variables as spatial maps (images). We now introduce the above two research directions.

1.2 Multimodal Learning

Multimodal learning refers to methods that can utilise and combine information from different modalities. Processing multimodal data poses several challenges, due to the heterogeneous information that the data encode [22]. Most commonly, a shared representation is sought, such that common and unique information is represented in the same latent space [23]. In this shared space, information from the different modalities is further similarly represented as modality invariant features, to allow fusion techniques [24].

In multimodal (as well as unimodal) image analysis tasks, such as synthesis and segmentation, fully convolutional networks are used to facilitate learning of spatial correlations between input and output. Thus, the shared representation is also spatial in the form of multi-channel feature maps. This thesis studies challenges of multimodal learning with fully convolutional networks,

and aims to produce modality invariant representations. This can be achieved by explicit biases, such as the ones imposed when learning disentangled representations, in which the modality-specific characteristics are disentangled from the remaining image features.

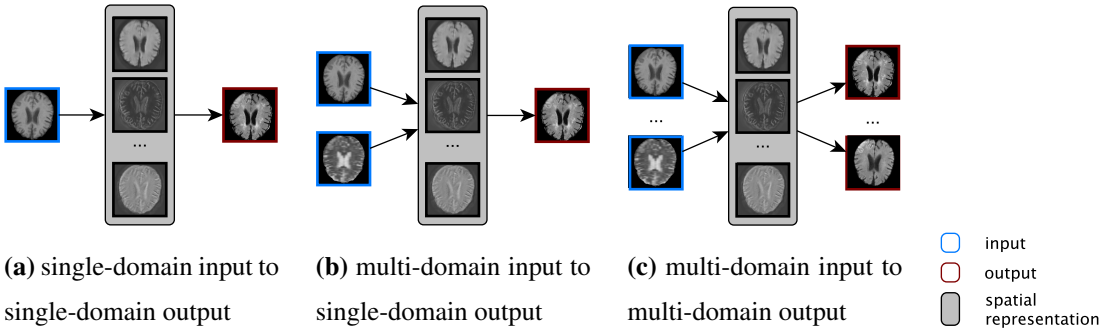
1.3 Disentangled Representations

Disentangled representation learning is a recent area of image analysis with deep learning that focuses on discovering the data generating factors. Such representations consist of factors, with each one corresponding to specific characteristics of the data distribution, such as image intensity, object orientation, size etc., and are widely used in many applications, for example in image translation [25] and pose estimation [26]. Since all input information is retained, these factors are not only useful for a particular task, but they can easily be extended to other related tasks [17]. Furthermore, disentangled representation learning is performed in an unsupervised way, and this is useful for tasks lacking annotations, such as in semi-supervised and transfer learning. Finally, since individual (or groups of) factors have some meaningful correspondence between specific image aspects, they promote model interpretability.

Nevertheless, learning disentangled representations remains challenging. Factors of variation are often not independent, can be of different dimensionality, and depend on inductive biases of the data and model design [27]. Specifically in both medical and computer vision context, images are disentangled in factors of structure and geometry (anatomy) and factors of appearance (image modality). The anatomical factors are of particular interest, since they are spatially represented, maintain pixel-wise correlations with the input and are thus useful in medical tasks.

1.4 Overview and Technical Contributions

We now give an overview of the thesis contributions. Considering spatial representations that are multimodal and disentangled, we aim to learn cardiac segmentation networks with less annotations that also benefit from different modalities. All proposed methods can be considered as encoder-decoders with the encoder mapping images to intermediate spatial representations and the decoder mapping these representations back to the image space. The following paragraphs briefly describe the proposed approaches, which investigate properties of *modality invariance*, *fusion*, and *disentangling factors*, for multimodal and semi-supervised learning. For the defini-



(a) single-domain input to single-domain output

(b) multi-domain input to single-domain output

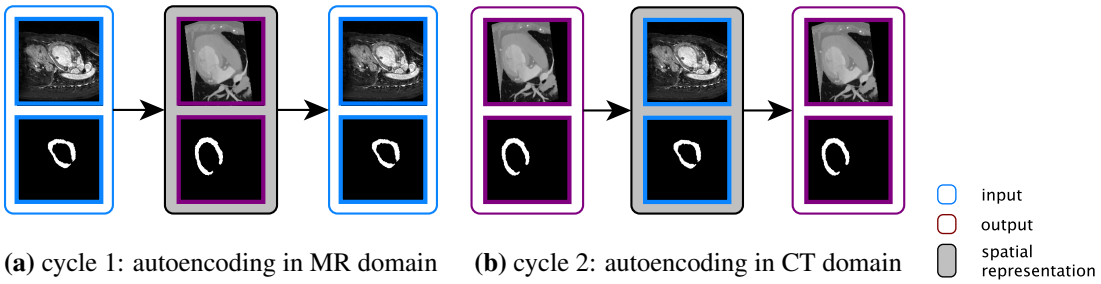
(c) multi-domain input to multi-domain output

Figure 1.3: Three examples of brain image synthesis with an encoder-decoder model and an intermediate spatial representation.

tions below, we consider sets of images and corresponding masks in input or output domains.

In Chapter 4 I propose a multi-input, multi-output fully convolutional network that encodes images of multiple domains to modality invariant feature maps. A schematic is shown in Figure 1.3, that corresponds to a model designed for MR synthesis of brain images in output domains from images in input domains, with domains here corresponding to multi-parametric MR modalities. I choose to first examine multimodal brain synthesis to avoid the more challenging cardiac images that are unregistered and non-isotropic, as described in Section 1.1. This model required all images to be spatially aligned, and showed that an intermediate spatial representation is capable of encoding multi-domain correlations. Depending on the availability of input domain data, and also the requirement for synthetic output domain data, a single model can use the same latent representation to perform predictions as follows: single-input to single-output (Figure 1.3a), multi-input to single-output (Figure 1.3b) or multi-input to multi-output (Figure 1.3c). Learning constraints on the spatial representation encouraged a modality invariant space, that is suitable for fusion techniques, in order to further combine features coming from the different inputs. This chapter's contribution is a new method for learning modality invariant representations, showing that spatial features represented as images of the same size as the input are suitable for combining multimodal information and improve synthesis quality. This model is published in two articles, in MICCAI 2017 with title "Robust multi-modal MR image synthesis" [28] and in IEEE Transactions on Medical Imaging in 2018 with title "Multimodal MR synthesis via modality-invariant latent representation" [29].

Chapter 5 extends this research for unpaired data, i.e. when there is no correspondence between images from the two domains. Here domains refer to different cardiac imaging techniques,

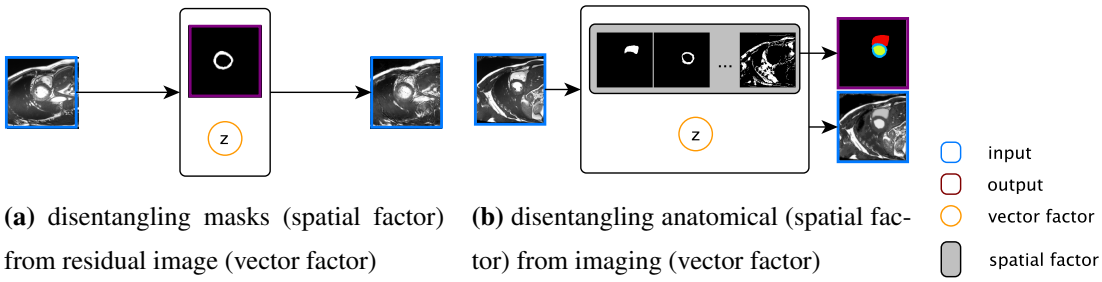


(a) cycle 1: autoencoding in MR domain **(b)** cycle 2: autoencoding in CT domain

Figure 1.4: Two cycles of cardiac image translation between MR and CT domains.

specifically MR and CT. The aim is to use the multi-domain data for cross-domain synthesis, since no information fusion is possible. The cycle consistency principle is adopted, which consists of two cycles and can also be seen as two autoencoders (one per domain). This is illustrated in Figure 1.4, where the first cycle learns mappings from MR to CT and back to MR, and the second cycle learns mappings from CT to MR and back to CT. Cycle consistency is typically used in domain translation when supervised learning cannot be applied. Although there is no latent space per se, a spatial representation can be considered as the output CT in the translation of the first cycle (Figure 1.4a), and respectively the output MR in the second cycle (Figure 1.4b). Furthermore, in order to guarantee spatial equivariance at each translation step, segmentation masks are concatenated with their respective MR and CT images. This chapter demonstrates that cross-domain cardiac synthesis is possible using unpaired data, and proposes a simple method for constraining translation functions showing the benefit of synthetic data as a data augmentation approach. This method is published in SASHIMI workshop of MICCAI 2017 with title "Adversarial image synthesis for unpaired multi-modal cardiac data" [30].

The cycle consistency principle is useful for translating between image domains, even when one domain is a semantic map. Indeed segmentation can be considered as a specialised form of image translation, however, the information content between the image and segmentation domains is different. Chapter 6 investigates translations between these two domains, and demonstrates the one-to-many problem when translating from a categorical to an image domain, since a semantic segmentation may correspond to many images. This problem is solved by encoding the residual information in a new vector variable, z, proposing disentangled representations in medical image analysis for the first time (Figure 1.5a). This model can be generalised to a semantic representation of the whole anatomy (spatial factor) as a multi-channel feature map with



(a) disentangling masks (spatial factor) from residual image (vector factor)

(b) disentangling anatomical (spatial factor) from imaging (vector factor)

Figure 1.5: Disentangled representations of spatial and vector factors.

the residual (vector factor) z, containing only imaging related statistics (Figure 1.5b).[1] The spatial factor is tightly correlated to both segmentation and other anatomical tasks. Additionally, autoencoding provides an unsupervised training mechanism. This chapter's major contribution is the first method for learning disentangled representations of anatomical and imaging features of medical images, as well as a detailed analysis of the properties and semantics of the latent factors. Furthermore, I demonstrate the use of such representations in semi-supervised and multi-task learning. This work has been published in two articles, the first in MICCAI 2018 titled "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation" [31] and the second in Medical Image Analysis in 2019, titled "Disentangled representation learning in cardiac image analysis" [32].

Inspired from the above, Chapter 7 presents a unified framework for multimodal and disentangled representations, illustrated in Figure 1.6. Multi-domain images are mapped to a disentangled representation of anatomical and imaging factors. Here domains refer to different cardiac MR modalities. Images although paired are not perfectly aligned, but the common spatial factor is suitable for correcting misalignments and therefore enables spatial fusion mechanisms. Here both autoencoding and cross-domain synthesis allow semi-supervised and even unsupervised learning when one domain has few or no mask annotations. This chapter's contributions consist of a new method that combines multimodal and disentangled representation learning to leverage information from multiple modalities for cardiac segmentation. Furthermore, disentangled representations offer robustness to input misalignments, to the amount of annotations and to multimodal image pairing. This method is presented in two articles, the first in STACOM workshop of MICCAI 2019, titled "Multimodal Cardiac Segmentation Using Disentangled Representa-

[1] In the literature the term "factor" usually refers to either a single dimension of a latent representation, or a meaningful aspect of the data (i.e. a group of dimensions) that can vary independently from other aspects. Here I use factor in the second sense to refer to a representation that consists of a (multi-dimensional) anatomy factor, and a (multi-dimensional) modality factor.

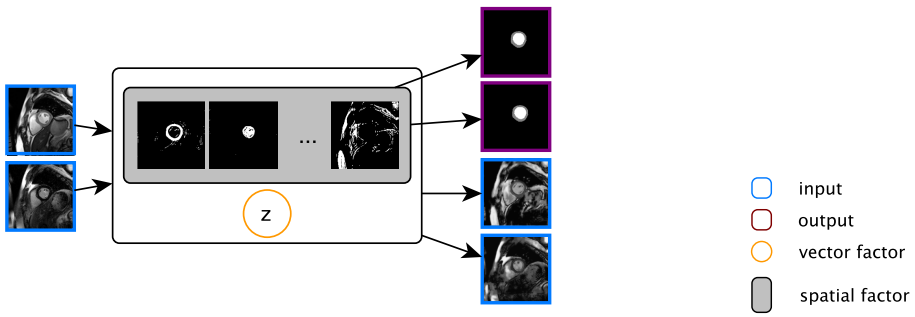


Figure 1.6: Multimodal and disentangled spatial and vector representations.

tion Learning" [33], and the second is under review in IEEE Transactions on Medical Imaging "Disentangle, align and fuse for multimodal and semi-supervised image segmentation" [34].

Finally, open source code of all proposed methods has been made publicly available to encourage dissemination also in other fields. Code is found under the following URLs:

- Chapter 4 - `github.com/agis85/multimodal_brain_synthesis`;
- Chapter 6 - `github.com/agis85/spatial_factorisation`, `github.com/agis85/anatomy_modality_decomposition`;
- Chapter 7 - `github.com/agis85/multimodal_segmentation`.

1.5 Clinical Significance

We propose methods for the analysis of brain and cardiac images and achieve contributions with potential clinical value. The proposed methods are based on deep learning and are thus data driven, meaning that they take advantage of available data and do not embed strong physiological priors. This can be advantageous for learning solutions on populations with common pathological conditions, although prior knowledge can be valuable to regularise and facilitate learning and also to enable specialisation in rare pathological conditions.

Methods of Chapters 4 and 5 offer an automated way of generating synthetic images by transforming images of the same subject in other modalities. This is most commonly used to enhance existing datasets with new images (data augmentation), or to replace images corrupted with artefacts (data imputation), for example due to motion. In Chapter 4 we propose a method that is able to increase the quality of synthetic images by jointly processing and combining

information in different MR sequences. Critically though, it does not require a specific number of input modalities, although benefits from multimodal inputs. This method is learned through pairs (or sets) of images of the same subject in multiple modalities that are perfectly aligned, common for example in multi-parametric brain MRI. However, this can be challenging when applied in cardiac image synthesis, for example between cine-MR and LGE modalities, since image acquisition is affected by cardiac motion, as well as by respiratory motion, which prevent obtaining perfectly aligned multimodal pairs. In Chapter 5, we relax the requirement for aligned multimodal pairs, and propose a method for cardiac synthesis that can use imaging data of different populations. By simultaneously transferring the myocardium annotations in the synthetic images, we demonstrate the importance of augmenting datasets with synthetic images when learning auxiliary tasks, such as when extracting myocardial segmentations.

Chapters 6 and 7 focus on cardiac segmentation in various MR modalities, where automatic methods are often challenged by the lack of large annotated datasets. We propose new methods that are robust to the number of annotations by employing semi-supervised, multi-task, and multimodal learning techniques. Specifically, we show that we can learn segmentation models with a fraction of images being annotated. Also, we can improve the model performance using auxiliary information from diverse sources if available, such as from the left ventricular volume. Finally, we show that our method can benefit from multimodal images, even if they are not perfectly aligned. In fact, processing with multiple inputs always yields improved segmentation performance, while also allows segmenting images from unannotated modalities.

1.6 Thesis Structure

Here we provide an overview of the thesis contents. Chapter 2 contains a background on medical imaging and presents the datasets used. Chapter 3 presents a technical background on deep learning and representation learning, and a literature review on the thesis research areas. Chapter 4 describes our method on multimodal brain MR synthesis. Chapter 5 discusses the cycle consistency principle for cardiac image synthesis and presents a synthesis application on data augmentation. Then, Chapter 6 proposes two new methods for disentangled representation learning in medical imaging, which respectively disentangle the myocardium and the anatomy. Chapter 7 aggregates our prior work on multimodal and disentangled representations to present a combined framework tested on various medical data and evaluate disentanglement. Finally, Chapter 8 concludes the manuscript, discussing limitations and future extensions of this work.

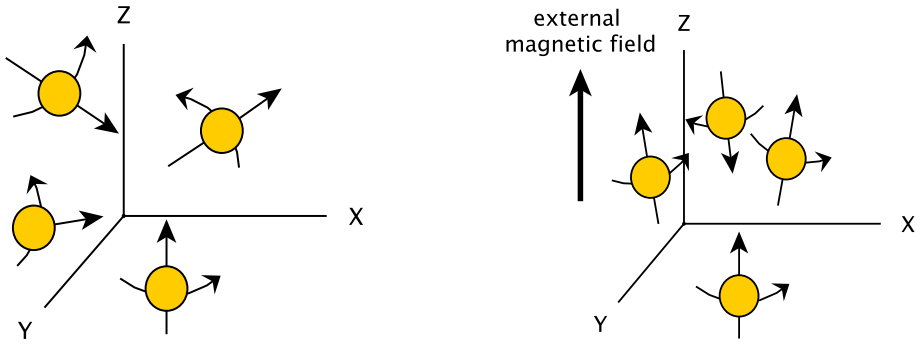
Chapter 2
Clinical and Medical Imaging Background

This thesis uses data from multiple imaging modalities, and specifically from Magnetic Resonance Imaging (MRI) and Computed Tomography (CT). We mainly focus on multi-parametric Magnetic Resonance Imaging (MRI), a non-invasive technique that uses magnetisation to image soft-tissues, and its application on cardiac image analysis. Although machine learning techniques do not typically take into consideration the physics of the image acquisition, some MR fundamentals are provided in Section 2.1, as well as an overview of CT in Section 2.2. Finally, a background on the physiology and functionality of the heart is presented in Section 2.4, as well as specific cardiac MR sequences in Section 2.5.

2.1 Magnetic Resonance Imaging (MRI)

MRI is extensively used for pathology detection in many organs, such as brain ischemia and cancer, abdomen lesions and tumours, and cardiomyopathies. Different parameterisation of the MR scanner generates sequences, termed modalities, that create contrast between adjacent organs and pathologies using different pixel intensities. Taking advantage of the fact that 70% of cells consist of water, MRI relies on the magnetisation of hydrogen atoms to visualise soft tissues in the body. Hydrogen atoms consist of single protons that are positively charged and have a spin, in other words they rotate around an axis at a constant rate. Under no external magnetic field, the direction of their magnetic field is random in space, as shown in Figure 2.1.a. The aim of MR imaging is to perform excitation followed by relaxation of these protons.

A MR scanner is a big magnet, that initially applies a magnetic field, which aligns hydrogen protons to the direction of the field, called the longitudinal direction, as shown in Figure 2.1.b. The proton spins are also partly aligned to the longitudinal direction, and spin at a frequency called Larmor frequency that depends on the strength of the magnetic field.



(a) no magnetic field

(b) external magnetic field on Z direction

Figure 2.1: Illustration of protons' magnetisation. (a) Under no external magnetic field, the protons have random directions. (b) When an external field is applied, the protons align to the direction of the field.

In order to align the proton spins, the MR scanner applies a radio frequency (RF) pulse with frequency equal to the Larmor frequency. This excitation phase aligns all proton spins, and changes their magnetisation direction to the transversal direction, which is typically perpendicular to the original field, but can also vary depending on the RF pulse. Receiver coils are then used to capture the energy emitted by the change in the magnetisation energy.

The excitation phase is followed by the relaxation phase, in which the RF pulse is stopped. This results in the protons changing their magnetisation direction to the original magnetic field, releasing energy. The time needed to achieve 63% of the original magnetisation is called T1 relaxation time. Stopping the RF pulse, also results in dephasing of the protons in the transverse direction, in which their spins are not aligned anymore. The time needed to dephase 37% of the original protons is called T2 relaxation time. T1 and T2 times differ between tissues because of their different concentration in water and fat. The energy released during the relaxation is used for image reconstruction. In fact, the final image is produced by the Fourier transform of the k-space image, an example of which is shown in Figure 2.2. This is a 2D array with dimensions equal to the image dimension, that consists of (k_x, k_y) points containing the phase and spatial frequencies of the image pixels [6]. Often additional external magnetic fields (gradients) are added, which in combination with the RF pulse constitute a MR sequence. MR images are predominately used in this thesis to evaluate synthesis (Chapters 4 and 5), and segmentation (Chapters 6 and 7).

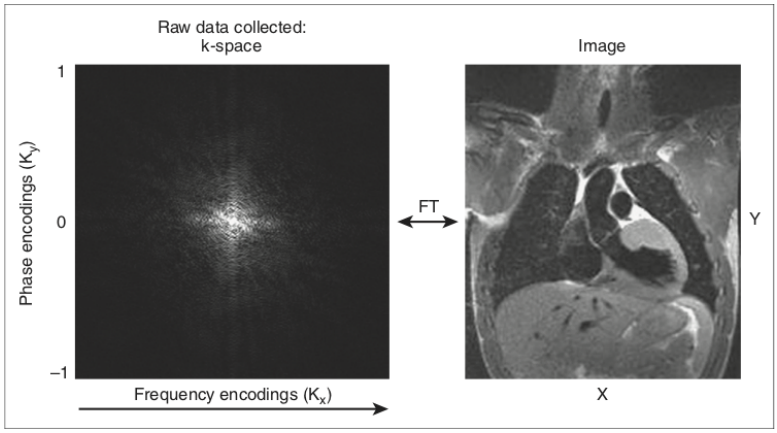


Figure 2.2: Reconstruction example from k-space. Image taken from [6].

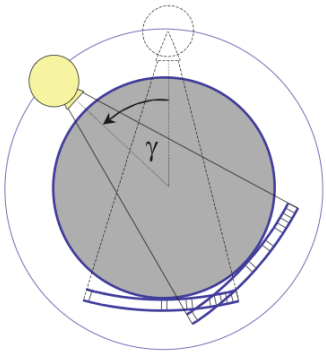


Figure 2.3: Schematic of a CT scanner, with the source emitting a conical beam from multiple directions. Here, γ represents the angle between two measurements. Image taken from [7].

2.2 Computed Tomography (CT)

CT measures the absorption of X-Ray radiation from the body tissues. Unlike MR, CT is a quantitative technique and measures the density of an organ by calculating the attenuation of X-Ray radiation. This is achieved with a detector that records the amount of X-Ray photons that are passed through the tissue of interest. The density is measured in Hounsfield units that depend on the composition of each tissue. Typically, as shown in Figure 2.3, a X-Ray source emits a cone-shape beam through different directions to calculate a 3D reconstruction using the backprojection technique [7]. An example CT image is shown in Figure 2.4. CT images are used for cross-modality synthesis to MR in Chapter 5.

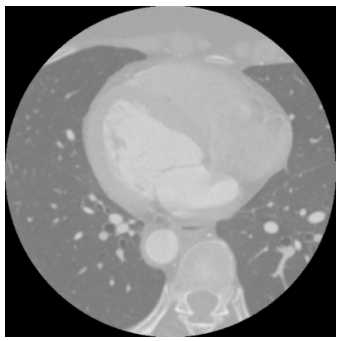


Figure 2.4: An example of a CT cardiac image. Image is taken from the Multimodal Whole Heart Segmentation (MMWHS) dataset [3,4].

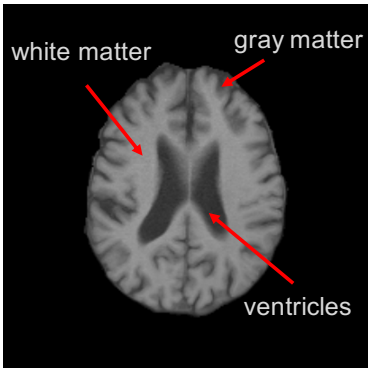


Figure 2.5: Example of a brain in T1w sequence. Gray matter, white matter and ventricles are marked with red arrows. Image is taken from Ischemic Stroke Lesion Segmentation (ISLES) dataset [1].

2.3 The Brain

We now give a brief description of important brain structures and of brain MR sequences that are used in Chapter 4. The largest part of the brain is the cerebrum that is divided into two hemispheres. The outer layer of each cerebral hemisphere is the cortex that is central to cognitive activity and is comprised of gray matter. The inner layer of the cerebrum consists of white matter and affects learning. Towards the centre of the brain, there is the ventricle system, which contains four ventricles that produce the cerebrospinal fluid. An example brain MR image with marked gray matter, white matter and ventricles is shown in Figure 2.5.

Common MR sequences for brain imaging that are also used in this thesis are T1 weighted (T1w), T1 contrast (T1c), T2 weighted (T2w), FLAIR, DWI, and PD, with some examples shown in Figure 2.6. T1w images are primarily used for healthy anatomy and consider the dif-

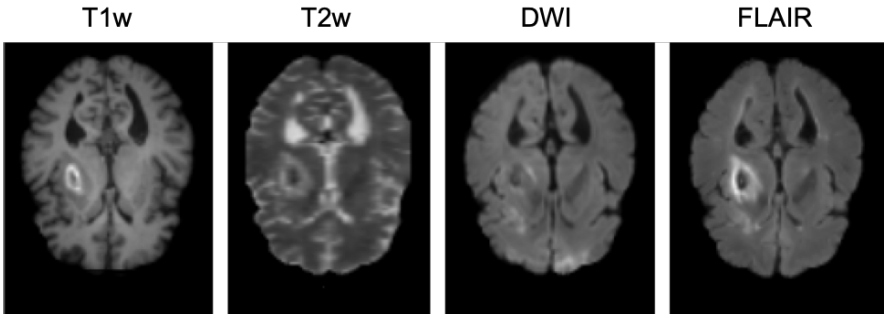


Figure 2.6: Brain images in T1w, T2w, DWI and FLAIR sequences. Depending on the intrinsic properties of each sequence, water and fat molecules are represented with different pixel intensities. Images are taken from Ischemic Stroke Lesion Segmentation (ISLES) dataset [1].

ferences in T1 relaxation time between fat and water with fat presenting higher pixel intensities in the reconstruction. T1 weighted images after the administration of a paramagnetic contrast agent, such as Gadolinium, produce T1c images that are often used to detect brain tumour. Similarly, T2w images consider the fat and water differences in T2 relaxation time. Here, water presents higher pixel intensities, and since it is correlated with edema, T2 images can detect pathologies. On the contrary, Proton Density weighted (PD) images do not consider neither T1 nor T2 signal, and rely on the number of protons in the image. Diffusion Weighted Images (DWI) are T2 images that measure the movement of hydrogen protons when fields of different magnetic strength are applied. They are used for detecting edema, for example in ischemia. Finally, Fluid Attenuated Inversion Recovery (FLAIR) is a T1w sequence that nulls fluids, such as the cerebrospinal fluid in brain, and can also detect pathologies. Multiparametric brain MR images are used for multimodal synthesis in Chapter 4.

2.4 The Heart

In the remaining chapters, we focus on the analysis of the heart that is comprised of several substructures, such as the ventricles, the atria, and the myocardium as shown in Figure 2.7. The heart is responsible for circulating oxygenated blood throughout the body using the myocardium (MYO) through the left atrium (LA), left ventricle (LV) and aorta respectively. It also circulates non-oxygenated blood to the lungs through the right atrium (RA), right ventricle (RV) and pulmonary artery [35]. This process happens throughout a cardiac cycle from end systole, when the myocardium is contracted, to end diastole, when the myocardium is ex-

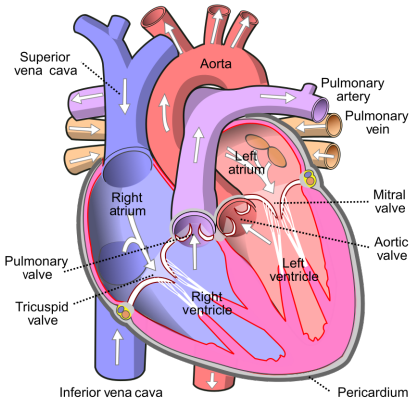


Figure 2.7: Substructures of the human heart with arrows indicating the blood flow. Image is taken from [8].

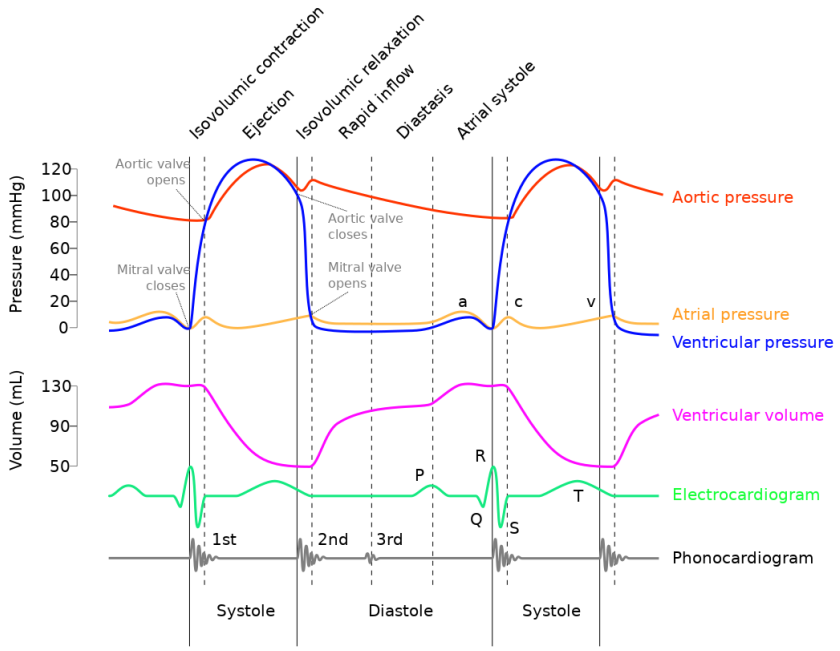


Figure 2.8: An example ECG showing the electrical activity of the heart, with the systole and diastole phases marked. Image is taken from [9].

panded. The heart's contraction and expansion is triggered by electrical signals that stimulate the myocardium and create a perfectly rhythmic cycle or heartbeat. This electrical activity is measured with an electrocardiogram (ECG), as shown in Figure 2.8, using electrodes on the skin. ECG is also used for imaging of the cardiac cycle, and is described in Section 2.5.

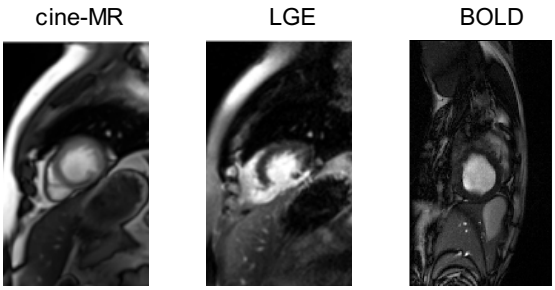


Figure 2.9: Examples of cardiac MR images in cine-MR, LGE, and BOLD modalities. Cine-MR and LGE images are acquired as part of the study in [2], and BOLD in [10]

2.5 Cardiac Imaging

Many imaging techniques exist for cardiac analysis, but here we focus on some protocols of Cardiac Magnetic Resonance (CMR), which are developed due to the high soft tissue contrast that MR exhibits, and because of its non-ionising property, as previously discussed in Section 2.1. Specifically, Sections 2.5.1, 2.5.2, and 2.5.3, describe cine-MR, Late Gadolinium Enhancement (LGE), and Blood Oxygen Level Dependent (BOLD) respectively, with representative examples shown in Figure 2.9. These are used in Chapters 6, and 7 for learning segmentation models.

2.5.1 Cine-MR

The most common CMR protocol is cine-MR, a temporal sequence typically consisting of 10-30 frames of the cardiac cycle that is used for calculating functional indices, such as the ejection fraction. It is also referenced as a bright-blood technique, since it generates high signal intensity for pixels within vessels compared to other tissues.

In order to achieve image acquisition of high quality, the k-space data for each frame are acquired across different cycles. The synchronisation of the sampled data to particular frames is performed through ECG gating that detects an R-wave which corresponds to the beginning of the systolic phase. MR imaging, and an ECG pulse defining the R-R interval of a heartbeat (see Figure 2.8), are run in parallel and synchronisation is performed retrospectively. An imaging session is completed within multiple breath-holds, and the scanning time for each cine-MR slice takes approximately 10 seconds. To reduce scanning time, non-isotropic images are taken with a low spatial resolution and slices typically between $8mm$-$10mm$.

2.5.2 Late Gadolinium Enhancement (LGE)

Other CMR protocols, such as the LGE include gadolinium, a contrast agent, for the detection of myocardial infarction [36]. In infarcted myocardial regions, gadolinium can penetrate cell membranes and appears bright in the image. After magnetisation of the heart with the radiofrequency pulse, the different recovery times of infarcted and normal myocardial tissue after the gadolinium has been injected, results in bright and dark (known as myocardium nulling) regions respectively. Imaging is performed only on the diastolic phase, and is used for detection of heart failure.

2.5.3 Blood Oxygen Level Dependent (BOLD)

In order to avoid radioactive contrast agents, other CMR techniques, such as BOLD MRI, consider oxygen that is present in blood cells as an endogenous contrast agent. The magnetic properties of blood cells depend on the amount of oxygen in hemoglobin, with the difference being in the T2 relaxation times. BOLD detects the infarcted myocardium by the smaller amount of oxygen supplied by stenosed vessels, which creates an inhomogeneous magnetic field within the myocardium, and appears as hyperintense regions. Unlike LGE, BOLD images are acquired across the whole cardiac cycle and not only to end diastole. Although the intensity differences are smaller compared to exogenous contrasts, it has been demonstrated that the BOLD effect is present in the heart, and can also be used for detecting infarcted regions in the myocardium [10].

2.6 Datasets

In this thesis we use various public and private medical datasets in MR and CT modalities to validate the developed methodologies. An overview of the datasets is presented in Table 2.1 with their sizes varying between 10 to 100 subjects and approximately 400 to 24,000 images respectively. This size is considered small compared to computer vision datasets, such as ImageNet [37] that has approximately 14 million images, but is typical for medical datasets in which data acquisition and distribution is more challenging and entails ethical and privacy processes. For research purposes the size of the employed datasets is sufficient to evaluate the proposed algorithms, however, a large study size would be required to evaluate the use of commercial applications in a clinical setting [38]. We now present a description of these datasets for brain (Section 2.6.1), cardiac (Section 2.6.2), and abdominal images (Section 2.6.3).

Dataset	Organ	Modality	Subjects	Task	Chapter
ISLES	brain	MR (T1w, T2w, FLAIR, DWI)	28	synthesis	4
BRATS	brain	MR (T1w, T1c, T2w, FLAIR)	54	synthesis	4
IXI	brain	MR (T1, T2, PD)	28	synthesis	4
MM-WHS	cardiac	MR, CT	40	synthesis, segmentation	5, 6
ACDC	cardiac	cine-MR	100	segmentation	6
QMRI	cardiac	cine-MR	26	segmentation	6
BOLD	cardiac	cine-MR, BOLD	10	segmentation	6, 7
ERI	cardiac	cine-MR, LGE	28	segmentation	7
CHAOS	abdomen	MR (T1, T2)	20	segmentation	7

Table 2.1: Overview of the datasets used in this thesis categorised based on organ, modality, size, task performed, and chapter used.

2.6.1 Brain

The multimodal synthesis work of Chapter 4 uses multi-parametric MR brain data from three datasets, with details summarised below.

2.6.1.1 Ischemic Stroke Lesion Segmentation (ISLES) - public

Ischemic Stroke Lesion Segmentation (ISLES) [1] data consists of 28 pre-processed volumes that are imaged in T1w, T2w, FLAIR, and DWI sequences. The volumes have been skull-stripped and re-sampled to an isotropic spacing of $1mm^3$, and co-registered to the FLAIR sequences. All volumes belong to patients with sub-acute ischemic stroke lesions, and were made publically available as part of a lesion segmentation challenge for MICCAI 2015.

2.6.1.2 Brain Tumour Segmentation (BRATS) - public

Brain Tumour Segmentation (BRATS) [16] data consists of high and low grade glioma cases, from which we used the latter containing 54 volumes, imaged in T1w, T1c, T2w, and FLAIR, and are released with segmentation masks of tumours. Data are skull-striped, co-aligned, and interpolated to $1mm^3$ resolution. Data were made available as part of a brain tumour segmentation challenge for MICCAI 2015.

2.6.1.3 Information eXtraction from Images (IXI)

Information eXtraction from Images (IXI) [39] data contains co-registered T1, T2 and PD-weighted non-skull stripped images from 28 healthy subjects. Data were collected at three London hospitals, specifically Hammersmith Hospital with a Philips 3T scanner, Guy's Hospital with a Philips 1.5T scanner, and the Institute of Psychiatry with a General Electric 1.5T scanner.

2.6.2 Cardiac

We further use various cardiac datasets with details presented below.

2.6.2.1 Multimodal Whole Heart Segmentation (MM-WHS) - public

Images from the Multimodal Whole Heart Segmentation (MM-WHS) challenge of MICCAI 2017 are used in Chapter 5 for cross-modal synthesis, and in Chapter 6 for multimodal segmentation as well as modality transformation and estimation.

The MM-WHS dataset contains 40 anonymised volumes, of which 20 are cardiac CT/CTA and 20 are cardiac MRI, made available by the authors of [3, 4, 20]. The CT/CTA data were acquired at Shanghai Shuguang Hospital, China, using routine cardiac CT angiography. The slices were acquired in the axial view. The inplane resolution is about $0.78 \times 0.78mm$ and the average slice thickness is $1.60mm$. The MRI data were acquired at St. Thomas hospital and Royal Brompton Hospital, London, UK, using 3D balanced steady state free precession (b-SSFP) sequences, with about $2mm$ acquisition resolution at each direction and reconstructed (resampled) into about $1mm$. Data contain static 3D images, acquired at different time points relative to the systole and diastole. All the data has manual segmentation of the seven whole heart substructures. Specifically: (1) the left ventricle blood cavity (LV), (2) the right ventricle blood cavity (RV), (3) the left atrium blood cavity (LA), (4) the right atrium blood cavity (RA), (5) the myocardium (MYO), (6) the ascending aorta (AO), and (7) the pulmonary artery (PA).

2.6.2.2 Automatic Cardiac Diagnosis Challenge (ACDC) - public

Images from the ACDC challenge [5] of MICCAI 2017 are used in Chapter 6 for semi-supervised segmentation and for latent space arithmetics.

This dataset contains cine-MR images acquired in 1.5T and 3T MR scanners, with resolution between 1.22 and 1.68 $mm^2/pixel$ and a number of phases varying between 28 to 40 images per patient. There are images of 100 patients for which manual segmentations are provided for the left ventricular cavity (LV), the myocardium (MYO) and the right ventricle (RV), corresponding to the end systolic (ES) and end diastolic (ED) cardiac phases. In total there are 1,920 images with manual segmentations (from ED and ES) and 23,530 images with no segmentations (from the remaining cardiac phases).

2.6.2.3 Edinburgh Imaging Facility QMRI - private

In Chapter 6 we also use images from Edinburgh Imaging Facility QMRI for semi-supervised segmentation and multi-task learning.

The dataset is acquired with a 3T scanner, and contains cine-MR images of 26 healthy volunteers each with approximately 30 cardiac phases. The spatial resolution is 1.406 mm^2/pixels with a slice thickness of $6mm$, matrix size 256×256, a field of view $360mm \times 303.75mm$, and image size 256×208 pixels. Manual segmentations of the left ventricular cavity (LV) and the myocardium (MYO) are provided, corresponding to the ED cardiac phase. In total there are 241 images with manual segmentations (from ED) and 8,353 images with no segmentations (from the remaining cardiac phases).

2.6.2.4 BOLD - private

A multimodal dataset of cine-MR and CP-BOLD images is used in Chapter 6 for modality estimation, and in Chapter 7 for multimodal segmentation.

This dataset contains 2D images from 10 (mechanically ventilated) canines with an in-plane resolution of $1.25mm \times 1.25mm$ that were acquired at baseline and severe ischemia (inflicted as controllable stenosis of the left-anterior descending coronary artery (LAD)) on a 1.5T Espree (Siemens Healthineers) on the same instrumented canines [10]. Images are acquired at short axis view covering the mid-ventricle and using cine-MR and a flow and motion compensated CP-BOLD, where each sequence is applied one after the other in the protocol in separate breath-holds. The pixel resolution is 192×114. This dataset (not publicly available) is ideal to show complex spatio-temporal effects as it images the same animal with and without disease, using two almost identical sequences that only differ in that CP-BOLD modulates pixel intensity with

the level of oxygenation present in the tissue. In total there are 129 cine-MR and 264 CP-BOLD images with manual segmentations from all cardiac phases.

2.6.2.5 Edinburgh Royal Infirmary (ERI) - private

In Chapter 7 we use a cine-MR and LGE dataset for multimodal segmentation.

The ERI dataset contains images from 28 patients [2], and is acquired at Edinburgh Royal Infirmary, with spatial resolution 1.562mm^2/pixel, and slice thickness $9mm$. End diastolic myocardial contours are provided. The image size is 192×192 pixels. The number of segmented images is 358 (for each of cine-MR, LGE).

2.6.3 Abdominal

2.6.3.1 Combined Healthy Abdominal Organ Segmentation (CHAOS) - public

Finally, multimodal segmentation of Chapter 7 is further evaluated in a T1-dual inphase and T2-SPIR abdominal dataset.

Combined Healthy Abdominal Organ Segmentation (CHAOS) [40] contains data released for the abdominal segmentation challenge [41, 42] that was part of ISBI 2019. Images of 20 subjects with liver, kidneys and spleen segmentations are acquired by a 1.5T Philips MRI scanner in T1-dual inphase and T2-SPIR sequences from PACS of DEU Hospital. In total there are 1594, 12-bit DICOM images of 256×256 resolution.

2.7 Overview

This chapter has presented background material on medical imaging techniques including MRI and CT, and discussed different MR sequences common in brain and cardiac imaging, as well as some physiological information on the anatomy of the heart. Finally, a description of the data used throughout the thesis is provided. The following chapter (Chapter 3) further expands the required background with technical preliminaries and definitions, as well as a literature review on synthesis and segmentation with recent deep learning methods.

Chapter 3
Technical Background

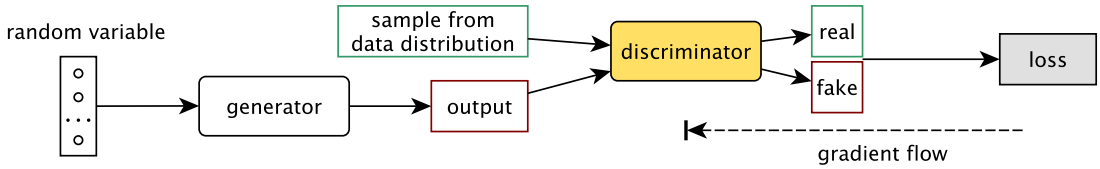
This chapter presents the technical background necessary of this thesis. Although a basic understanding of machine learning and deep learning principles is assumed, a brief introduction to learning with various degrees of supervision is provided in Section 3.1. Then, Sections 3.2 and 3.3 describe in detail two widely used generative models, namely Generative Adversarial Networks (GAN) [43], and Variational Autoencoders (VAE) [44, 45], which are extensively used throughout this thesis. Finally, Section 3.4 presents the benefits of representation learning.

3.1 Model Learning

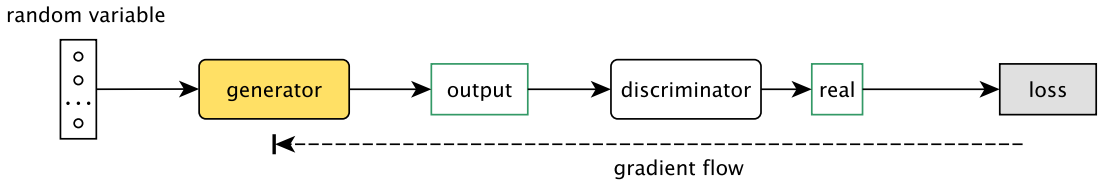
Most commonly in machine learning we assume a dataset of N pairs of datapoints $\{(x, y)\}_1^N$, where x is a sample from an input distribution $x \in X$, and y is a sample from an output distribution $y \in Y$, and the task is to learn a mapping function, $f : X \rightarrow Y$. Learning such functions can be performed with neural networks, which are proven to be universal function approximators [46]. Neural networks consist of layers of hidden units that extract features from the input data, such that they can predict the target output data. The learning process with networks of multiple layers is termed *deep learning*.

Learning a function between input and output data defines discriminative models and can be characterised as supervised, semi-supervised, and unsupervised depending on the relative size of the two sets of data, and the pairing of input and output samples. On the one hand, if each input sample x has a corresponding output sample y, then learning is considered supervised. On the other hand, if there are more input than output samples, then learning is termed as semi-supervised. Finally, unsupervised learning concerns cases where there are no pairs of input and output samples, and the function is learned only based on prior beliefs.

Naturally given enough paired data samples, supervised learning is usually more accurate. However, and as previously discussed in Chapter 1, this can be problematic in medical image analysis, where data acquisition is expensive. Therefore, this motivates research for semi-



(a) Training step 1: discriminator training (generator weights are not updated)



(b) Training step 2: generator training (discriminator weights are not updated)

Figure 3.1: Schematic of a Generative Adversarial Network (GAN). A generator transforms a random sample from a known distribution to an output sample. The discriminator classifies samples from the real distribution and outputs from the generator. Training is performed in two steps with gradients updating the weights of the discriminator or the generator, respectively: (a) the discriminator is trained to classify real and synthetic samples; (b) the generator is trained to produce outputs that are classified as real by the discriminator.

supervised and unsupervised methods. One way to learning with no supervision is with generative models that capture the data generating processes, or in other words the causal relationships between data and generating factors. Two popular generative models are discussed in the next two sections, GANs and VAEs. They are based on adversarial training and variational inference respectively, and have had successes in many tasks, such as in image generation [47]. Here we use GANs in the cross-modal synthesis method of Chapter 5, and the VAE formulations when learning disentangled representations in Chapters 6 and 7.

3.2 Generative Adversarial Networks (GAN)

The aim of generative models is to approximate a probability distribution function, typically using maximum likelihood estimation, although this is intractable for unknown data distributions. A GAN [43] is a framework for approximating some data distribution using two networks: a generator G and a discriminator D. The generator learns a mapping function from a known distribution, e.g. a Gaussian, to a target distribution, and the discriminator classifies samples between the true and predicted distribution. The two networks are trained adversarially, such

that the generator maximises the discriminator's loss, an analogous to a min-max game. A simple schematic can be seen in Figure 3.1.

More formally, given a data distribution $p_{data}(x)$ and a prior $p_z(z)$ of a random variable z that typically follows a multivariate Gaussian $p_z(z) \sim \mathcal{N}(0, I)$, the loss function is the following:

$$L(G, D) = \min_G \max_D \mathop{\mathbb{E}}_{x \sim p_{data}(x)} [\log(D(x))] + \mathop{\mathbb{E}}_{z \sim p_z(z)} [\log(1 - D(G(z)))]. \tag{3.1}$$

Training is performed with stochastic gradient descent, by alternating gradient updates of the parameters of D and G. Specifically, D is trained to correctly classify samples from the real or the predicted distribution, and G is trained adversarially to maximise D's classification loss. Convergence is achieved when G can generate a probability distribution equal to the data distribution, i.e. when $p_g = p_{data}$.

At convergence, GANs learn a smooth function, where nearby input values correspond to similar synthetic samples. The smoothness of the learned function can been qualitatively evaluated with synthetic images produced by interpolating the input vector [48]. However, as shown in [49], the generated distribution is biased by the support of the real distribution, which affects the "steerability" of GANs and their ability to generalise beyond the training data.

Nevertheless, adversarial training of deep neural networks is challenging, and sensitive to many variables. Common issues include mode collapse, training instabilities, and vanishing gradients. In mode collapse the generated distribution consists of a part or a mode of the true distribution, and occurs because there is no explicit cost for diversity. Training is unstable if the Generator and Discriminator oscillate rather than converging to a fixed point. Finally, vanishing gradient problems occur if the rate of convergence is different and one agent becomes more powerful than the other.

A lot of research has focused on improving training stability, for example with careful network architecture design [48]. Furthermore, different losses have been proposed to replace the original binary cross entropy. In [50], a least-squares loss is shown to promote a smooth and non-saturating gradient. The Generator minimises the squared distance between synthetic examples and the decision boundary, thus heavily penalising points that are classified away from the boundary. Moreover, the Wasserstein loss has been proposed to measure the distance between the real and generated distribution [51,52]. In this formulation the Discriminator minimises the earth-mover distance, i.e. the cost of mass needed to move from one distribution to

the other. This has been shown to overcome the problem of vanishing gradients produced when the generated distribution is far away from the real distribution. Another popular approach for smoother training minimises the largest singular value of Discriminator's weight matrices [53], and can be applied to any GAN architecture. Furthermore, in BEGAN [54], the Discriminator is modelled as an autoencoder, and the real and generated images are first autoencoded before comparison with the Wasserstein distance. This prevents the Discriminator from easily becoming a good classifier.

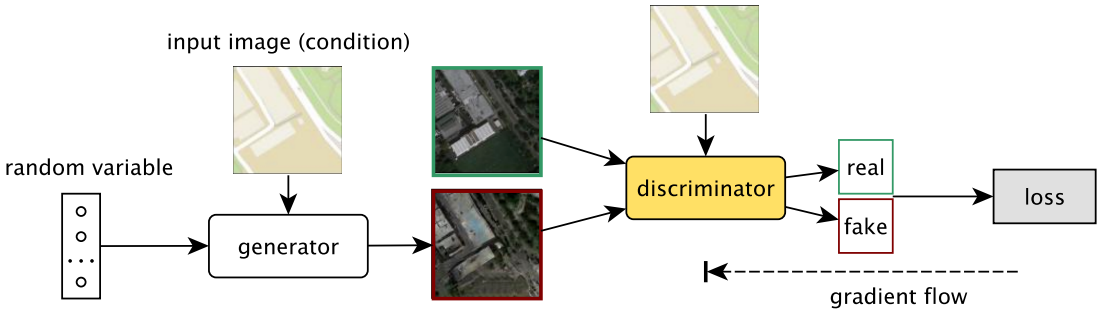
Many approaches have also been proposed to alleviate the mode collapse problem. Batch normalisation has been suggested to help the gradient flow in deep models, and also force some variation within samples from each batch, which can make collapse less likely [48]. VEE-GAN [55] learns a function mapping the real data to a Gaussian, and then uses this to encourage the generated data to also result in a Gaussian distribution when put through the same function. This provides a training signal to the generator that comes from outside of the discriminator.

An evaluation of popular GAN architectures has been performed in [56], and showed the sensitivity of adversarial methods to random initialisation, hyperparameters and datasets. In addition, they showed that given enough computational power and good hyperparameters, comparable results can be achieved by most architectures. Extensive reviews of different GAN variants have also been performed in [47, 57].

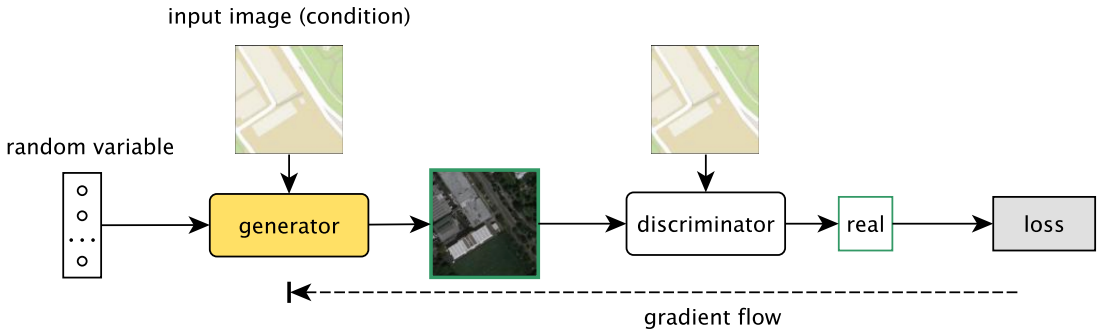
Except for generative modelling, adversarial training has also offered a new type of loss function, which relaxes the need for input-output pairs that is required by traditional discriminative machine learning. Since the discriminator classifies real from synthetic examples, it learns a prior over the real distribution, and thus an unsupervised loss can be implicitly defined, which can be used in combination with supervised losses to constrain the space of predicted outputs. This will be demonstrated in Chapters 6 and 7, where an adversarial unsupervised loss constrains the shape of the segmentation masks.

3.2.1 Conditional GANs

In the context of this thesis, the conditional GAN variant [58] is relevant, in which the adversarial generative model is conditioned by a variable. When both the condition variable and the output are spatial, i.e. images, GANs learn spatial mapping functions. Typical examples include image to image translation or image synthesis, the task of transforming images between



(a) Training step 1: discriminator training (generator weights are not updated)



(b) Training step 2: generator training (discriminator weights are not updated)

Figure 3.2: Schematic of an image-conditional GAN for image-to-image translation between two domains. Given a random sample and an image in the input domain (condition variable), the generator produces an output image of the same content in an output domain. The input image also conditions the discriminator. Training follows the classic GAN formulation as follows: (a) the discriminator is trained to classify real and synthetic samples; (b) the generator is trained to produce outputs that are classified as real by the discriminator. Images show map to photograph synthesis and are taken from [11].

two domains. Figure 3.2 shows an image-conditional GAN for synthesis of photographs from maps, in which the discriminator is trained to classify real from fake pairs of maps and photographs, and the generator is trained to predict realistic photographs. A popular architecture is Pix2Pix [11], in which a neural network receives two inputs, a random sample from the Normal distribution, and an input image in one domain (condition variable), and is trained to predict an output image to a second domain. The network is trained with a supervised cost using real target images and an unsupervised cost using adversarial training, where a discriminator classifies predicted images as real or fake, i.e. belonging to the distribution of the target domain or not.

A type of conditional GAN, CycleGAN [59], is used in Chapter 5, when we translate images

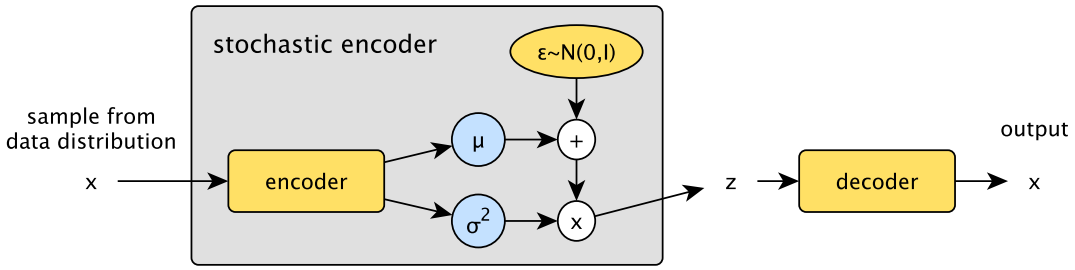


Figure 3.3: Schematic of a Variational Autoencoder (VAE). A stochastic encoder maps an input sample to a probability distribution with a mean and variance. The decoder, using the re-parameterisation trick, draws a sample from the predicted distribution to reproduce the input.

between MR and CT. CycleGANs overcome the lack of paired data by using a cycle consistency loss, in which each image is first translated to the other domain and then reconstructed. Furthermore, and as mentioned in Section 3.2, conditional GANs are also used as shape priors for unsupervised segmentation in Chapters 6 and 7.

3.3 Variational Autoencoders (VAE)

A different view to approximating a data distribution is using the VAE framework [44, 45]. This consists of two networks, a decoder or generator, and an encoder or inference model. The decoder maps samples from a prior latent distribution, $z \sim p(z)$, to samples of the data distribution, $x \sim p(x|z)$, whereas the inference model maps samples from the data distribution to the latent variables, $z \sim p(z|x)$. A VAE schematic is displayed in Figure 3.3. In summary, a VAE maintains the auto-encoding principle, i.e. that of reconstructing the input, with the difference that it assigns a probability distribution on the latent space, typically a multivariate Gaussian, to every input sample.

VAEs are probabilistic models, and also latent variable models assuming that data samples x are generated by sampling a likelihood $x \sim p(x|z)$ from unknown latent factors $z \sim p(z)$. The aim is to evaluate or infer the posterior distribution of the latent variables given the observed data $p(z|x)$. Using the Bayes rule, this posterior is $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$, where $p(x)$ is obtained by marginalising the latent variables, $p(x) = \int p(x|z)p(z)dz$. However, calculating an analytic solution for the posterior or the integral of the marginal distribution is intractable, for example because of high-dimensional data or because of complex forms of distributions. In such cases, variational methods are employed to approximate the posterior $p(z|x)$ with a

parametric distribution $q(z|x)$. In the VAE model, the choice of $q(z|x)$ is a multivariate Gaussian, the parameters of which (mean and variance) are embedded in the estimation of the latent variable z by the stochastic encoder. This distribution represents the range of plausible z values that correspond to a sample x.

Training the VAE is performed with maximum likelihood estimation by maximising the marginal log-likelihood $\log p(x)$, which is defined as follows:

$$\begin{aligned}\log p(x) &= \mathop{\mathbb{E}}_{q(z|x)}[\log \frac{p(x,z)}{p(z|x)}] = \mathop{\mathbb{E}}_{q(z|x)}[\log \frac{p(x,z)}{q(z|x)} \frac{q(x,z)}{p(z|x)}] \\ &= \mathop{\mathbb{E}}_{q(z|x)}[\log p(x,z) - \log q(z|x)] + \mathop{\mathbb{E}}_{q(z|x)}[\log \frac{q(z|x)}{p(z|x)}] \\ &= L_{ELBO} + KL(q(z|x)||p(z|x)).\end{aligned} \tag{3.2}$$

According to equation 3.2, the marginal log-likelhood is equal to the sum of the Evidence Lower Bound (ELBO), $L_{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x,z) - \log q(z|x))$, and the Kullback Leibler (KL) divergence, $KL(q(z|x)||p(z|x))$. Since the KL divergence is non-negative, maximising the ELBO with respect to $q(z|x)$ concurrently minimises the KL and pushes the approximate probability $q(z|x)$ to match the true $p(z|x))$. In summary, the VAE loss function is written as:

$$\begin{aligned}L_{VAE} &= \mathop{\mathbb{E}}_{q(z|x)}[\log p(x,z) - \log q(z|x))] \\ &= \mathop{\mathbb{E}}_{q(z|x)}[\log p(x|z) + \log p(z) - \log q(z|x))] \\ &= \mathop{\mathbb{E}}_{q(z|x)}[\log p(x|z) - KL(\log q(z|x)\| \log p(z))].\end{aligned} \tag{3.3}$$

In practice, the likelihood $p(x|z)$ is modelled with a decoder neural network and corresponds to the reconstruction cost of a datapoint x, and the approximated $q(z|x)$ is modelled with an encoder network $f(x)$. However, training is problematic since it requires back-propagating the loss across random samples $z \sim q(z|x)$, where $q(z|x)$ depends on the parameters of the encoder. To solve this, the random variable z is reparameterised as a deterministic variable $z = f(\epsilon, x)$, and a random variable $\epsilon \sim p(\epsilon)$ is introduced that is independent of the model parameters. This reparameterisation trick allows the VAE model to be differentiable. In the case that $z \sim \mathcal{N}(\mu, \sigma^2 I)$ the encoder network outputs the parameters μ and σ for each input sample x, and the reparameterisation of z is equal to $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

In summary, VAE models embed the observed data in a smooth manifold of latent variables.

In order to minimise the KL-divergence (with a multivariate Gaussian), posterior distributions should have similar means and variances, which results in a high overlap between the distributions of each data point and on average decreases the discriminability of distributions. This may also increase reconstruction error, since a sample drawn from a distribution given one input may have higher probability to be drawn from the distribution of a different input. In such ambiguous cases the decoder will predict an output that is an average of data points. Therefore, in order to decrease reconstruction error, VAE training converges to encoding similar data inputs in close points in the latent space (smooth manifold). Furthermore at inference time, VAE models can be used as generative models by sampling from a Gaussian and decoding this sample. Unlike GANs, and since the data density is estimated, the decoded samples cover the data distribution and do not suffer from the problem of mode collapse. We take advantage of the smooth VAE manifolds in Chapters 6 and 7 to embed intensity distributions of medical images from single or multiple modalities respectively.

3.4 Representation Learning

Generative models, such as GANs and VAEs, learn a mapping from a structured latent representation, i.e. one that approximates a Gaussian distribution, to an unknown data distribution, i.e. that of images. Additionally, VAEs, and also Adversarial Autoencoders [60], learn a reverse mapping from the data to the representation. Representation learning is a long running goal of machine learning [17], with good representations being typically those that capture explanatory (discriminative) factors of the data, and are useful for the task(s) considered.

In supervised learning, deep networks generate such representations at every layer to maximise the posterior $p(y|x)$, i.e. the probability of task y given input x. Typically no restrictions on the features are applied, and they are thus useful for a particular task, but cannot generalise to other tasks. Often to accommodate this, and also to improve generalisability to unseen data, either models are trained for multiple related tasks, or explicit restrictions are applied, as in the VAE case, where features are samples from a known distribution (see Section 3.3). The latter assumes that there is one "bottleneck" representation layer, where all data can be mapped onto. This is useful to encourage richer representations through unsupervised learning, which not only describe the data, but are also useful for the task at hand, and thus enable semi-supervised learning. Most commonly, unsupervised learning of representations is achieved with autoencoders. Autoencoders also describe the methods described in this thesis.

3.4.1 Autoencoders

Autoencoders are deep neural networks that consist of two components, an encoder and a decoder, and are trained to reconstruct the input through an intermediate representation z. The encoder function f learns a mapping from the input to the latent features $z = f(x)$, and the decoding function learns a reverse mapping from the features back to the input space $y = g(z)$. The aim is to learn discriminative features, and therefore autoencoders should not learn to copy information, but rather learn mappings to and from a low dimensional manifold, such that nuisance factors are ignored. This is achieved by constraining the latent representation z to be of lower dimensionality compared to the input. Although, z is most commonly represented in a vectorised form, here we are interested in spatial representations. This renders autoencoders as fully convolutional, and as we will see in the following chapters, makes them useful for spatially equivariant tasks, such as synthesis and segmentation.

We have so far described an autoencoder's latent features as a compressed representation of the input, but simply compressing the input does not make a representation useful for learning tasks. One could simply consider a sufficiently complex encoder and decoder that could reconstruct the input through highly compressed features, but would not be useful in complementary tasks. An intriguing question that arises is: what properties should good latent representations have?

3.4.2 Defining Good Representations

Specific properties that characterise good representations are discussed in [17]. These include *smoothness*, which helps generalisation to test data that are encoded near points of the training data, and capturing *multiple explanatory factors*, which is a property of disentangled representations. Furthermore, a latent representation typically consists of the deepest layers of a network, produced by a series of *hierarchical* layers, and encodes more abstract features that are invariant to noise, and can be *shared across tasks*. As previously mentioned, representations enable *semi-supervised learning*, since unsupervised training estimates the joint distribution $p(x, y)$ of data x and task y, and thus the discovered features are also useful for the posterior $p(y|x)$. Moreover, representation learning is based on the *manifold* theory, in which the data probability mass lies in a low dimensional space, e.g. in autoencoders.

Learning good representations for medical imaging tasks poses several challenges, since often there is limited annotated data. Also the representation must lend itself to a range of useful

tasks, and work across data from various image modalities. These challenges can be approached by multimodal and disentangled representations.

3.4.3 Representations for Multiple Modalities

Multimodal representations embed common and unique information from complementary data sources. In order to encourage learning of cross-modal correlations, objectives such as cross-modal synthesis of data in one modality to the other, can be introduced [23]. Many methods further impose restrictions for making the multimodal representations similar [24, 61]. Such representations allow the joint consideration of multimodal data by fusing at different levels, classifying these techniques in early, middle and late fusion ones [62]. In multimodal imaging data, where convolutional neural networks are employed, the feature fusion requires their spatial alignment, something that is not always guaranteed. Since image registration can be challenging due to differences in intensity resolutions [63], alignment can be performed in the feature representation space.

However, alignment and fusion are applicable only when multimodal data are paired. The setting that consists of data of an annotated source and an unannotated target modality (or domain), where the multimodal data are not necessarily paired, is termed domain adaptation. A similar problem is domain generalisation, in which the data from the target domain are unseen during training. In both cases, the aim is to perform the same task on data of all domains, and is typically approached by learning domain invariant representations [64, 65].

In this thesis we extensively study multimodal representation challenges, focusing on fusion in Chapter 4, and on cross-modal synthesis in Chapter 5. We also approach multimodal registration challenges in Chapter 7 through disentangled representations, which offer the potential to separate the modality-specific characteristics.

3.4.4 Representations with Disentangled Factors

Disentangled representations capture information about input data in many (independent) factors, so that a change in the direction of one factor influences some meaningful aspect of the data [17] (hence also encountered as factorised representations). This ability promotes interpretability, for instance when changing an object's position and shape [44, 45]. Furthermore, learning is unsupervised and representations can be simultaneously useful in many tasks [17].

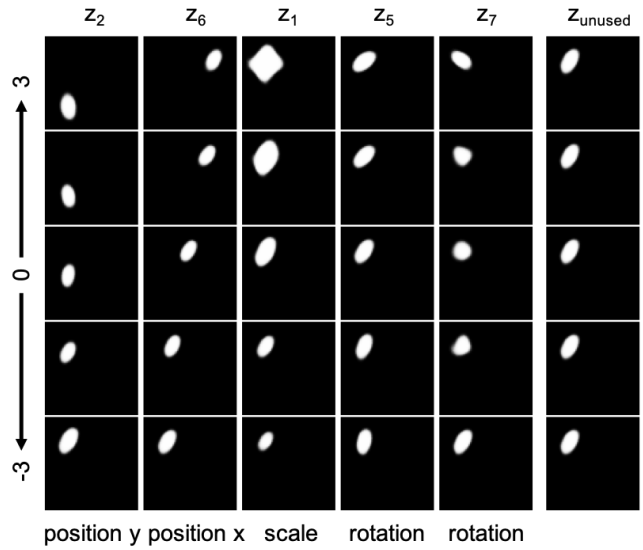


Figure 3.4: A disentangled representation learned by β-VAE on synthetic shapes. Each column shows synthetic images when interpolating a dimension z_i between -3 and 3. The effect of each dimension on the shape is indicated at the bottom. The final column corresponds to an unused $z-$dimension with all its values producing the same image. Image taken from [12].

Learning disentangled representations is made possible by the inherent properties of the VAE. As discussed in Section 3.3, the posterior $q(z|x)$ converges to a Gaussian prior $p(z) \sim \mathcal{N}(0, I)$ with a diagonal covariance matrix, which results in independent, i.e. factorised dimensions. According to [12], high factorisation is achieved when the factors of variation are aligned with the axes of the Gaussian posterior. To this end, it is important to increase the weight β of the KL-divergence in Equation 3.3, which however, constrains the capacity of the representation and, as described in Section 3.3, affects reconstruction quality. In order to improve reconstruction under a compressed representation, the $\beta-$VAE aligns the (conditionally independent) generative factors of the data, i.e. those with high contributions to reconstruction quality, to the dimensions of the Gaussian posterior. This preserves the smoothness of the representation and also achieves a disentangled representation. This trade-off between disentanglement and reconstruction is discussed in [12]. An example from [12] is shown in Figure 3.4.

Recent methods extend the notion of disentanglement in spatial and vectorised latent spaces. Although, strict independent constraints are not enforced, the visual quality is significantly improved for example when performing image translation [25], and the spatial representation is further useful for spatially equivariant tasks, as we will see in Chapters 6 and 7.

3.5 Literature Review

This section presents published literature that is linked to the methods proposed in this thesis. Methods for multimodal learning are discussed in Section 3.5.1, and a review of disentangled representations is presented in Section 3.5.2. Finally, Section 3.5.3 presents methods for cardiac segmentation with full and semi-supervision, as well as with multimodal learning.

3.5.1 Multimodal Learning

Multimodal machine learning is an active research area, as evidenced by recent methods on segmentation [66, 67] or classification [68] tasks. This is natural as multimodal images concern the same subject, but provide different information to be exploited. In this thesis we investigate multimodal image analysis in Chapters 4, 5 and 7 for synthesis and segmentation.

A recent taxonomy on multimodal learning [22] has identified the following challenges: *representation*, *translation*, *alignment*, *fusion*, and *co-learning*. Representation refers to learning informative features; translation refers to having the ability to transform data from one modality to the other; alignment refers to the process of learning relations between objects of each modality; fusion refers to the ability of joining unique information present only in one of the modalities; finally co-learning refers to the challenge of training a model such that knowledge from one modality helps the other. While in computer vision modalities might refer to any heterogeneous source of information, such as text and images, here, as common in the medical domain, we restrict to different image acquisitions, i.e. MR and CT.

In the following we focus on a common multimodal application, that of image synthesis. A background on unimodal synthesis with methods using a single modality is firstly discussed in Section 3.5.1.1. Then we present literature on multimodal synthesis using two or more modalities in Section 3.5.1.2. Section 3.5.1.3 discusses challenges of learning multimodal representations, and finally Section 3.5.1.4 describes a related problem in domain adaptation.

3.5.1.1 Unimodal Synthesis

MR synthesis has often been treated as a *patch-based* regression for example to produce pseudo-healthy data [69] and to synthesise CT from MRI [70]. In this setting mappings are learnt, using various techniques, which take a patch of an image or volume in one modality, and predict the

intensity of the central pixel of the corresponding patch in a target modality. The performance of these approaches has been shown to be aided by the addition of hand-crafted features that capture elements of the global structure of the image [71].

Another common approach to synthesis is the use of an *atlas*, such as in [72, 73]. Here, rather than learning a mapping, an atlas of image pairs is leveraged, and reconstructing a new volume from a source modality is achieved by matching the volume with the entries in the atlas of the same modality, and constructing the synthetic images from the corresponding atlas images in the target modality.

A sparse *dictionary* representation of the source and target modality has been proposed in [74], which synthesises new images with patch matching. In [75], joint dictionary learning is used to learn a cross-modality dictionary of the pair of source and target modalities that minimises the statistical distribution between them via optimisation. Image synthesis has also been treated directly as an *optimisation* problem in an unsupervised setting [76], where the target modality candidates are generated by a search method and then combined to obtain a synthetic image.

More recently, *neural networks* have been applied to MR synthesis and segmentation, and like many of the sparse coding based methods, often they approach the problem as a patch based regression [77]. The Location Sensitive Deep Network (LSDN) [78] is a patch-based neural network that, given as input a patch and its spatial position within the volume, can learn a position-dependent intensity map between two modalities. Motivated by the observation that conditioning on the location in the volume greatly reduces the complexity of the intensity transforms needing to be learnt, LSDN has been shown to produce state-of-the-art MR synthesis results. Another neural network approach is [79], in which a deep encoder-decoder network synthesises images of a target modality. Neural networks have also been employed to synthesise pseudo-healthy images. A GAN-based approach is proposed in [80], whereas in [81], a VAE synthesised pseudo-healthy images for the purpose of image registration.

In Chapter 4, we also adopt an encoder-decoder network for MR synthesis, but after this work, new methods employed adversarial networks to improve the image quality. For example GANs and perceptual losses are proposed for brain synthesis [82], and in MedGAN along with latent feature losses that regularise the style and content of the output [83]. Moreover, a 3D patch-based convolutional network for MR to CT synthesis [84] regularises synthesis with image gradients, and iteratively refines results. Image gradients have also been added using Sobel

filters in EA-GAN [85]. Wasserstein and perceptual losses have been proposed for CT image denoising [86]. Synthesis with WassersteinGAN has also been proposed for brain ageing [87].

The above methods assume paired input and output data, and thus cross-modal mappings are directly learned with supervision losses. Concurrently to our cardiac synthesis method of Chapter 5, CycleGANs have been used for unpaired synthesis of brain MR and CT [88], and for pelvic synthesis in combination with a gradient loss [89]. As we will see in Chapter 5, CycleGANs do not guarantee anatomy preservation during the translation process, and are prone to introducing geometric transformations. Therefore, they have been regularised with segmentation losses, as in [90]. A comprehensive review can be found in [91].

Cardiac Applications

There has been little previous work on learning-based methods for cardiac synthesis. Such methods have been explored for super-resolution, i.e. spatial up-sampling [92], and can be learned by creating a low resolution version of a dataset, and then learning to synthesise the original resolution, again admitting a supervised approach. Recently, cardiac super-resolution has been enhanced by incorporating a shape prior in the learning process [93]. Furthermore, super-resolution has been coupled with cross-modal synthesis using dictionary learning: with the addition of unpaired data in the learning process, a weakly supervised learning approach has been proposed [94].

A conditional GAN has been used for synthesising systolic from diastolic frames [95] to evaluate if different pathological conditions affect cardiac motion. Segmentation masks can be used as part of synthesis, as for example in our method of Chapter 5 to produce synthetic labelled images, or to regularise cardiac synthesis of MR to CT [90] or of cine-MR to LGE modalities [96]. Finally, recent work used disentangled representation frameworks to translate cine-MR to LGE for data augmentation [97], or for temporal synthesis of the cardiac cycle to regularise segmentation [98].

3.5.1.2 Multimodal Synthesis

Multimodal synthesis attempts to improve synthetic results by combining images, that are often spatially aligned. The first deep learning approach used a multi-input, multi-output encoder-decoder, and will be described in Chapter 4. Related methods include the single input, multi-

output method, Extended Modality Propagation [99]. Unlike related methods, where the input is expected to be an image in some source modality, in [99] the input is a label map, which delineates the areas of interest (e.g., white and grey matter), and the algorithm synthesises multimodal images accordingly. Using a dictionary of aligned multimodal images coupled with their respective label maps, extended modality propagation finds the most probable patch to apply at each location of the output image. However, it uses a single input and solves a somewhat different problem. Synthesis with multiple inputs has further been approached with random forest regression in image patches of multiple scales [71]. Since our work of Chapter 4, multimodal synthesis has been extended with many published methods, as described below.

GANs in combination with an auto-context mechanism is used to transform a fused image in the modality of interest [100]. This fused image is produced by a locally adaptive fusion, which uses a 3D convolutional kernel to weigh each encoded input modality. Concatenated multimodal images for multi-input, multi-output synthesis were used in MM-GAN [101] with channels of zeros indicating the missing/target modality. A similar method that uses cycle consistency losses proposed CollaGAN when there are no paired data of the target modality [102]. Concatenated inputs with cycle consistency were also proposed in DiamondGAN [103] but for multi-input, single-output synthesis, and using a binary vector of ones and zeros indicating which input modalities are available. Training single and multiple streams in MustGAN, in combination with a joint network for information fusion of features at multiple layers showed superior MR synthesis performance [104].

3.5.1.3 Multimodal Representations

Perhaps one of the reasons that multimodal synthesis has been difficult to accomplish is the need to map data into a common shared representation, such that it maintains the properties of Section 3.4. In Chapters 4 and 7 we investigate spatial representations that are suitable for multimodal learning, while focusing on robustness to missing data, as well as in data fusion.

Previous work on multimodal data fusion and shared representation in neural networks [105] has shown the plausibility of shared latent representations for generative tasks. There has also been relevant work on common representation learning, in which different data types are embedded into a common representation space. Key early work on multimodal learning that was robust to missing data is the multimodal autoencoder [23], in which a bimodal deep autoencoder was learnt for audio and video speech data. This model could reconstruct both modalities from

either the audio or the video, and was trained by minimising this reconstruction error. However, as noted in [61], there is no direct learning signal encouraging a shared common representation. In an attempt to address these shortcomings Correlational Neural Networks [61] both directly encourage correlation in the common representation space, and minimise the cross reconstruction error. A similar approach was also proposed in [106], where two autoencoders with tied weights are trained to learn the mapping between modalities. Obtaining similar latent features has been proposed with minimising a cosine similarity loss [107], or the KL-divergence when using the VAE model [108]. Correlational Neural Networks have been extended for sequential data with the use of Recurrent Networks [109]. However, their current formulation restricts them to the bi-modal setting, due to the use of explicit correlation calculations. In addition a statistical regularisation approach is proposed in [24], in which cross-modal scene representations are learnt, and the regularisation is done by encouraging the latent representation activations for all modalities to follow the same distribution. Similarly to the above, we encourage similar multimodal representations in Chapter 4 by minimising their variance, and in Chapter 7 by specific spatial constraints that make the latent features binary.

More recently, multimodal representations that are similar, are encouraged with adversarial training. This offers flexibility in that no explicit distance metric is required for the respective correlations. Cross-modal GANs [110] utilise two discriminators for this purpose. The first discriminates between features coming from either modality against the encoders that aim to generate common features, in order to learn a shared representation between images and text. The second discriminates between features from a real or synthetic sample to improve synthetic quality. The shared space is also encouraged with weight sharing of the encoders.

3.5.1.4 Representations for Domain Adaptation

Related to multimodal learning is domain adaptation, where the aim is to learn a representation through unpaired unimodal inputs. Usually the different domains consist of images in different appearances, but with similar structures, and annotations are provided for one domain. Typically multi-domain images are mapped to a common representation, that is used for particular task. An example is Domain Adversarial Neural Networks [64], which learn domain invariant features with adversarial training, and use these features as an input to a task classifier. A discriminator is trained to classify features coming from either domain, and a gradient reversal layer is used to achieve the domain invariance. Different domain representations with com-

mon and unique information are proposed in Domain Separation Networks [111]. Separating the unique information is achieved by an orthogonality constraint between the representation matrices of each domain, whereas the common information are encoded in domain invariant features: these are learned with adversarial training as in [112], and also by minimising the Maximum Mean Discrepancy [113] between the features.

Furthermore, autoencoders are also common in domain adaptation. In [114] a dual-input autoencoder is applied on images and cartoon sketches for image translation to produce cartoon sketches. Learning involves a similarity loss on the bottleneck, specifically the Mean Squared Error (MSE) between an encoded real and synthetic image. Synthetic quality is improved in XGAN [115] that uses specialised encoders and decoders for each domain, and applies two losses on the representation: an adversarial loss for domain invariance as in [112], and a semantic consistency loss as in [114]. A similar methodology for unsupervised image translation with a shared representation uses multiple VAEs [116], with the domain invariance achieved with weight sharing of layers near the VAE bottleneck, as well as with the Gaussian constraint.

Domain or modality invariance is another aim of this thesis, particularly investigated in Chapters 4 and 7. Furthermore, multimodal representations contain relevant information for a particular task, that is accumulated from data of various modalities. However, which information is relevant is strongly task dependent. A different approach would be to decompose the input into meaningful components, thus creating a disentangled representation, and will be shown in Chapter 6. This maintains all information about the data, separated in distinct factors.

3.5.2 Disentangled Representations

Interest in learning independent factors of variation of data distributions is growing. To date, methods have focused on representing factors of variation as independent latent variables, using Autoencoders [117] or VAEs [21] to decompose classification related factors from remaining image reconstruction factors. VAE [44, 45] were used for unsupervised learning of factorised representations, where the factors of variation are discovered throughout the learning process [12, 118]. For example β-VAE [12] adds a hyperparameter β to the KL-divergence constraint, whilst Factor-VAE [118] boosts disentanglement by encouraging independence between the marginal distributions. Furthermore, InfoGAN was proposed in [119], in which mutual information between a latent variable and the generated images is maximised. More recently, feature decompositions were proposed for video data to separate foreground from

background [120], and motion from content [121]. SD-GAN [122] generates images with a common identity and varying style. Combinations of VAE and GANs have also been proposed, for example by [123] and [124]. Both learn two continuous factors: one dataset specific factor, in their case class labels, and one factor for the remaining information. To promote independence of the factors and prevent a degenerate condition where the decoder uses only one of the two factors, mixing techniques have also been proposed [125]. These ideas also begin to see use in medical image analysis: 3D VAEs are applied to learn a latent space of cardiac segmentations that would be useful for disease diagnosis [126]. Learning factorised features is also used to distinguish between (learned) features specific to a modality and those shared across modalities [127]. However, their aim is combining information from multimodal images and not learning semantically meaningful representations.

The above methods learn factorised representations in the form of latent vectors. However, spatial information could be directly represented in a convolutional map, and this would benefit spatially equivariant tasks, such as segmentation. Typically this entails a disentanglement of style and content, where content is the spatial factor and style the non-spatial factor. This is also central in the disentanglement method of Chapter 6 for semi-supervised segmentation.

3.5.2.1 Style and Content Disentanglement

Disentangling style from content for style transfer is gaining popularity in computer vision, with many examples such as the seminal work of [128]. Classic style transfer methods do not explicitly model the style of the output image and therefore suffer from style ambiguity, where many outputs correspond to the same style. In order to address this "many to one" problem, a number of models have recently appeared that include an additional latent variable capturing image style. For example, colouring a sketch may result in different images (depending on the colours chosen) thus, in addition to the sketch itself, a vector parameterising the colour choices is also given as input [129].

Many disentanglement models use the vector and spatial representations for the style and content respectively [25, 130, 131]. Augmented CycleGANs [130] extend the CycleGAN framework for translating between domains with loss of information, for instance between images and semantic maps. An additional variable captures this additional information and transforms the one-to-many mapping to many-to-many. Multimodal unsupervised image to image translation [25] extends [116] that learned a multimodal embedding with domain specific encoders

and decoders. In [25] the representation is separated in a convolutional content and a vector style, such that samples from the style distribution in combination with a particular content can be decoded to the domain of interest. A similar decomposition and training regime is also used in [131] for image to image translation. Furthermore, [132] expressed content as a shape estimation (using an edge extractor and a pose estimator) and combined it with style obtained from a VAE. Their extension [26] also learns an encoder from images to a part-based content and a style vector using careful design choices. Disentangled representations have also been used for image deblurring [133], where the two factors are the content and blur features. Deblurring is achieved by decoding the content without the blur features.

Semantic representations have recently been pursued in computer vision in the form of feature masks [134] or by learning geometry with landmarks [135]. In [136] images are separated in landmarks and a style vector, although their re-entanglement first transforms the landmarks back to vector space, thus losing the spatial semantics.

We differentiate from the above techniques by proposing in Chapter 6 spatial representations modelled as categorical feature maps, in order to achieve both semantic and quantifiable properties, such that they can be used by multiple relevant tasks.

3.5.2.2 Disentangled Representations in Medical Image Analysis

In medical imaging, the disentanglement of content and style most often corresponds to the disentanglement of anatomical and imaging information. For example, such disentangled representations have found application in registration [137], where the registration field is calculated in the modality invariant anatomy instead of the image space. Furthermore in multi-task classification of ultrasound images by disentangling anatomical from shadow artefacts with adversarial learning [138] or in disentangling domain from category features using mutual information [139]. Disentanglement of liver lesion type regarding texture and contrast, and lesion shape can be used to synthesise images of new lesions for data augmentation purposes [140]. Moreover, disentanglement of metal artefacts in CT from the anatomy can be used for synthesising artefact-free images [141]. Different types of disentanglement, for example that of lung nodule from the background has been proposed for lung nodule synthesis [142].

Disentangled representations enable jointly processing multiple modalities, for example in image translation tasks, such as retina synthesis [143]. The modality invariant content space is also

suitable for segmentation tasks with domain adaptation. Multiple modalities have been used for liver segmentation with domain adaptation [144, 145], albeit without information fusion, and also for brain tumour segmentation [146], using though registered images. Finally, metal artefact disentanglement in CT modalities is jointly trained with segmentation outputs [147].

As discussed here, a common application of disentangled representations is on segmentation. In fact, disentanglement can further be applied to semi-supervised learning [21]. This is important for medical image analysis tasks, which often suffer from a lack of annotated data. Chapters 6 and 7 propose methods to deal with this problem by using unannotated data for semi-supervised cardiac segmentation, and by utilising information from a secondary modality, respectively. Below we present related literature on cardiac, as well as on semi-supervised segmentation.

3.5.3 Medical Segmentation

Cardiovascular diseases are ranked first worldwide in mortality causes, with 17.9 million reported cases per year according to World Health Organisation [148]. Accurate segmentation of the cardiac substructures, as well as temporal analysis of the heart has a great diagnostic value, because of the functional indices that can be calculated. Examples include the ejection fraction for diagnosing heart failure, which is calculated as the LV (or RV) volume difference between the end diastole and systole divided by the diastolic volume [149], and the LV wall thickness for diagnosing hypertrophic cardiomyopathy [150].

3.5.3.1 Supervised Cardiac Segmentation

Automatic cardiac segmentation has been studied extensively, especially in cine-MR, with particular focus on accurately delineating the myocardium and ventricles, as ventricular volume is useful for patient assessment and diagnosis. Previously, competitive methods were based on atlases [4], deformable models, level sets and machine learning methods [20]. Then, neural networks were employed for region of interest cropping [151] or contour initialisation [152], with level sets performing segmentation.

Current state-of-the-art segmentation results are obtained using deep convolutional neural networks, trained on labelled data, that directly learn segmentation masks from images [153, 154] or treat segmentation as a regression task in polar co-ordinates [155]. Correlations between adjacent slices of the cardiac volume have been modelled with recurrent neural networks [156].

Cascaded networks [157] are used to perform 2D segmentation by transforming the data into a canonical orientation and also by combining information from different views. Many methods have been proposed, see for example the participants of workshop challenges [5]. In specific cases, results can match human evaluation as in the study of [158].

Prior information on the cardiac shape has been used to improve segmentation with explicit conditioning [159] or by mapping predictions to a manifold of segmentation masks [93]. Moreover, temporal information related to the cardiac motion has been used to segment all cardiac phases [160, 161]. Also, segmentation has been regularised with multi-view images [162].

Although, most methods are applied to 2D slices, extensions to 3D also exist. A fully convolutional nework with supervision is proposed in [163], whereas [164] explore spatial correlations between adjacent slices to consistently segment 3D volumes. For a recent review see [165].

A common problem in medical image segmentation methods is the lack of big annotated datasets. Manual segmentations are a laborious task, particularly as inter-rater variation means multiple labels are required to reach a consensus, and images labelled by multiple experts are very limited. This can be challenging when training deep learning models. Thus typically augmentation techniques with spatial, intensity, and elastic deformations or with synthetic data (see Chapter 5) are employed. Using image translation or domain adaptation as augmentation strategy to reconcile lack of annotated data has been proposed with cycle consistency [90, 96, 166, 167], and disentanglement [97] losses. In domain adaptation, multimodal images are related with different augmentations [168], histogram matching [169] or adversarial losses [170]. A different line of approach takes advantage of the (usually) large number of unannotated images with semi-supervised learning.

3.5.3.2 Semi-supervised Segmentation

Semi-supervised segmentation has been proposed for cardiac image analysis using an iterative approach, where a convolutional network is alternately trained on labelled and post-processed unlabelled sets [171]. GANs were used in [172], for a gland segmentation task, involving supervised and unsupervised adversarial costs. Another approach [173] aims to minimise the distance between embeddings of labelled and unlabelled examples by comparing them in feature space. More recent medical semi-supervised image segmentation approaches include [174] and [175]. In [174] they address a multi-instance segmentation task in which they have bound-

ing boxes for all instances, but pixel-level segmentation masks for only some instances. [175] approach semi-supervised segmentation with adversarial learning and a confidence network. More recently, an additional classification task of the area of the left ventricle has been used as a regulariser for the segmentation task, using both labelled and unlabelled data [176]. Regularisation by predicting anatomical positions, that are already available, has also been shown to improve segmentation when training with a small number of subjects [177].

Semi-supervised learning with GANs was also proposed for semantic segmentation. The discriminator classifies between real and synthetic segmentation masks produced by the generator in [178], while in [179] the generator is used to increase the dataset size and the discriminator performs segmentation. Semi-supervised segmentation through data augmentation can also be achieved by learning deformation and intensity transformations of the labelled data by using the unlabelled images [180]. For a recent review see [181].

Although many categories of semi-supervised learning have been proposed, for example using iterative and adversarial losses, or by encouraging shared features, our methods of Chapter 6 take advantage of reconstruction costs for utilising unannotated images. The lack of annotations can be alternatively addressed if the same subject is imaged in another modality. Multimodal learning does not only help with lack of annotated data, but also when segmenting one modality is more challenging, for instance due to reduced tissue contrast. Here we differentiate in two categories of multimodal learning with registered and unregistered images.

3.5.3.3 Multimodal Segmentation with Registered Images

Early work on multimodal deep learning concatenated co-registered multimodal images in different input channels, in order to improve MR brain segmentation [18]. Common feature representations were achieved with multiple encoders that were proposed for cross-modal classification [182]. Furthermore, generalising to new unseen modalities has been studied in [127] using feature factorisation into modality descriptive and modality conditioned features.

Another aspect of multimodal learning is information fusion, used to combine complementary information. Most commonly, fusion is performed on the latent features [18], although, fusion at multiple levels can be achieved with densely connected layers [183, 184] to exploit multi-scale correlations. The fusion operator can also be learned with a fully connected layer [185] to merge multimodal temporal information describing disease progression. Furthermore, cross-

modal convolutions are used as a way to weigh each modality's contribution [186]. Finally, attention modules and residual learning focus on specific regions for brain MRI segmentation [187]. Chapters 4, and 7 adopt the max-fusion operator for combining multimodal information, which has the advantage of selecting the most informative features, and does not depend on specific number of inputs.

3.5.3.4 Multimodal Segmentation with Unregistered Images

Image misalignments are common in multimodal data. In the brain, registration can be reliable but in the heart and other moving organs performance cannot be guaranteed. Correcting misalignments at feature space is possible with Spatial Transformer Networks (STN) [188], which can be incorporated as layers in the training process, see Chapters 4 and 7. In summary, STNs make registration a differentiable operation, such that it can be part of a neural network. A STN consists of three components, a localisation network, a grid generator and a sampler. The localisation network, given a source and target feature map, predicts the parameters of a transformation, for instance an affine transformation matrix. The grid generator applies the learned (differentiable) transformation on a set of points arranged in a regular grid. Finally, the transformed feature map is the result of a differentiable sampler that applies a sampling kernel, e.g. bilinear interpolation, on the transformed grid and target feature map.

An alternative to feature alignment is with encoder-decoder setups that can learn shared features by co-learning with multimodal data. An exploration of different setups [189] showed that separate encoders and decoders sharing the last and first layer achieve the highest performance.

In cardiac image analysis approaches are limited. Multiple inputs can be combined by adapting segmentation masks with contour models [190, 191]. Alternatively, reducing the field of view to the patch level and ensembling using results from several atlases can alleviate the effect of committed errors [4]. A recent work [192] proposes simultaneous segmentation and registration of multimodal CMR by modelling the joint distribution with Multivariate Mixture Models.

Multimodal images can further be used as different samples of the same data distribution to form an expanded dataset [193], or be used for fine tuning of a pre-trained network [194]. Masks of the same subject from a labelled modality can be "proxy" for an unlabelled modality with a carefully balanced segmentation loss [195]. Finally, multimodal registration, although susceptible to errors, can create "noisy" labels [196], or concatenated multimodal pairs [197].

Our approach in Chapter 7 combines disentangled representations with STN to correct misaligned multimodal features, and perform multimodal fusion on aligned features.

3.6 Metrics

The methods of the subsequent chapters are evaluated by comparing with state-of-the-art benchmarks. Given 2D images $x, \hat{x} \in X$ and segmentation masks $m, \hat{m} \in M$, where x and m are real and $\hat{x}$ and $\hat{m}$ are predictions from some model, and $X \subset \mathbb{R}^{H \times W}$, $M \coloneqq \{0,1\}^{H \times W \times L}$ with H, W being the image height and width respectively, we consider the following metrics.

Mean Squared Error (MSE). MSE is the mean squared difference and is defined as:

$$MSE(\hat{x}, x) = \frac{1}{H * W} \sum_{h \in H} \sum_{w \in W} \left[\hat{x}(h,w) - x(h,w)\right]^2. \tag{3.4}$$

Mean Absolute Error (MAE). MAE is mean absolute difference and defined as:

$$MAE(\hat{x}, x) = \frac{1}{H * W} \sum_{h \in H} \sum_{w \in W} \left|\hat{x}(h,w) - x(h,w)\right|. \tag{3.5}$$

Structural Similarity Index (SSIM). SSIM measures image quality. Given μ_x and σ_x^2 the mean and variance of image x, and $\sigma_{\hat{x}x}$ the covariance between x and the prediction $\hat{x}$, SSIM is computed as follows:

$$SSIM(\hat{x}, x) = \frac{(2\mu_{\hat{x}}\mu_x + c_1)(2\sigma_{\hat{x}x} + c_2)}{(\mu_{\hat{x}}^2 + \mu_x^2 + c_1)(\sigma_{\hat{x}}^2 + \sigma_x^2 + c_2)}. \tag{3.6}$$

Peak Signal to Noise Ratio (PSNR). PSNR also is a measure of image quality and is computed as follows, where MAX_x is the maximum pixel value of the image:

$$PSNR(\hat{x}, x) = 10 \log_{10} \left(\frac{MAX_x^2}{MSE(\hat{x}, x)} \right). \tag{3.7}$$

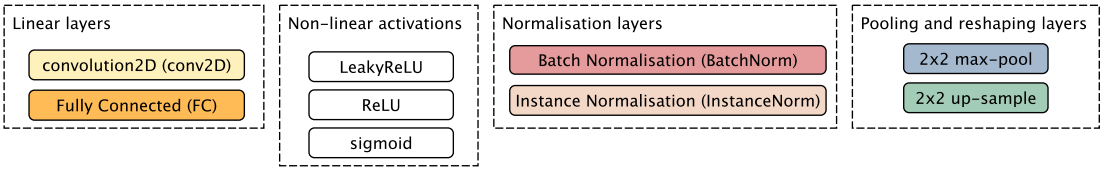


Figure 3.5: Graphical representation of different neural network layers. These are categorised in layers defining a linear operation (convolutional and fully connected layers), non-linear functions (Leaky ReLU, ReLU, sigmoid), normalisation layers (batch and instance normalisation), and pooling and reshaping layers (max-pooling and nearest neighbour up-sampling). When applicable, the number of layer parameters are also provided, e.g. $3 \times 3 \times 64$ $conv2D$ defines a 2D convolution with 3×3 kernels resulting in a 64-channel feature map, and $FC50$ a fully connected layer with 50 neurons.

Dice coefficient. Dice is a measure of overlap, and is used to evaluate categorical images, e.g. segmentation masks. Dice is defined as:

$$DICE(\hat{m}, m) = 2\frac{|\hat{m} \cap m|}{|\hat{m}| + |m|}. \tag{3.8}$$

3.7 Model architecture graphs

All proposed methods are based on neural networks, which are defined as a series of linear and non-linear computational blocks. The network architectures can therefore be described graphically using a glossary of components that are defined in Figure 3.5. These components are used in the figures of the following chapters to depict the architectures of the proposed methods, and can be grouped in categories depending on the performed operation. These categories are linear layers, that include 2D convolution and fully connected layers, non-linear activation functions, such as the Rectified Linear Unit (ReLU), Leaky ReLU and sigmoid functions, normalisation layers at batch or instance level, and finally downsampling and upsampling operations that include max pooling and 2D upsampling with nearest neighbour layers.

3.8 Overview

This chapter has discussed background material on technical deep learning preliminaries. We have presented literature on multimodal and disentangled representation learning, described

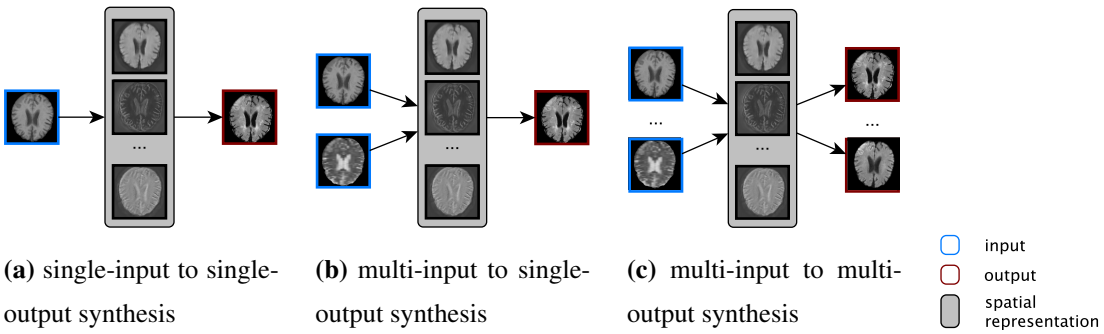
challenges, and also methods from the medical and computer vision communities. Finally, we have focused on cardiac segmentation, a task challenged by the lack of annotations, and have presented related literature on supervised, semi-supervised, and multimodal segmentation. The following chapters present our approaches to multimodal and disentangled representation learning that are evaluated on medical synthesis and segmentation tasks. Note that each following chapter contains a related work section to highlight the state-of-the-art literature at the time that the methods and corresponding publications were released.

Chapter 4
Multimodal Image Synthesis

4.1 Introduction

In this chapter we investigate representations for multimodal learning. We show that images, specifically multi-channel feature maps produced by neural network encoders, are suitable latent variables to capture and combine spatial information from multiple inputs. As briefly described in Chapter 1, we approach the multimodal learning challenge through brain image synthesis tasks shown in Figure 4.1, and investigate different modality synthesis setups of single-input to single-output (Figure 4.1a), multi-input to single-output (Figure 4.1b), and multi-input to multi-output (Figure 4.1c).

By synthesis here we mean a model that takes a number of images as input, showing the same organs in different modalities, and outputs synthetic images of that same anatomy in one or more new modalities. Image synthesis [198] has attracted a lot of attention due to exciting



(a) single-input to single-output synthesis

(b) multi-input to single-output synthesis

(c) multi-input to multi-output synthesis

Figure 4.1: Three examples of brain image synthesis with different number of input and outut modalities and an intermediate spatial representation.

This chapter is based on:

- Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A., 2018. Multimodal MR synthesis via modality-invariant latent representation. IEEE Transactions on Medical Imaging, 37(3), pp. 803-814.
- Joyce, T., Chartsias, A., Tsaftaris, S.A., 2017. Robust multi-modal MR image synthesis. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 347-355). Springer, Cham.

potential applications in medical imaging: synthesised data for example may be used to impute missing or corrupt images [199], to derive images lacking a particular pathology, which is not present in the input modality [69, 80], to improve algorithm performance on other medical imaging tasks, such as image segmentation [199], and others.

The current state-of-the-art methods in image synthesis learn mappings between pairs of image modalities [71, 76, 78]. However, it is often the case that we have several modalities available (a typical clinical MR protocol collects a multitude of images), and taking advantage of their collective information could potentially improve synthetic results. In fact, different modalities highlight different anatomy (or pathology) in the body and, by using them together, it is possible to obtain better synthesis results through information sharing. For this reason, state-of-the-art methods use multi-input architectures [71] and obtain higher quality synthetic images. On the other hand, if a specific number of input modalities is *mandatory* for a model, then this reduces the number of applicable cases to the ones strictly containing this complete set of image modalities. To overcome this we propose a multi-input (and multi-output) deep neural network, which does not require all inputs in order to synthesise outputs, but can make use of additional inputs, when available, to achieve enhanced accuracy.

An additional consideration when dealing with multiple inputs, is the misalignment between the images of different modalities. The inclusion of a Spatial Transformer Network (STN) [188] makes our model more robust to such misalignments. Another key problem in MRI synthesis is that many different MR scanners are used, and the different images produced (of the non parametric type) have non-identical statistical properties, which typically require several pre-processing steps to alleviate. Thus, an algorithm trained on images from a particular scanner may degrade significantly in performance when applied to images from other sources. To address this, we demonstrate transfer learning by fine-tuning a trained decoder with a very small number of volumes from a different source.

4.1.1 Approach Overview

The proposed end-to-end model, illustrated in Figure 4.2, takes 2D images as input, making use of multiple modalities when available, thus allowing users to simply provide any of the available modalities at test time. It outperforms a neural network, and random forest method when trained on a single modality, with results improving further when additional modalities are given as input. The model processes input images in four stages: encoding, alignment,

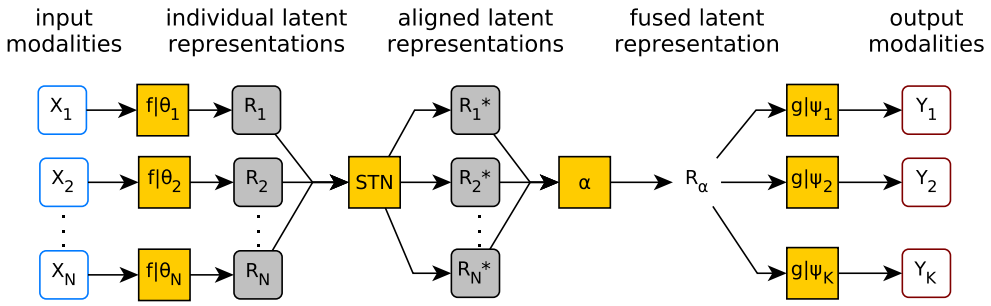


Figure 4.2: Model schematic for multimodal synthesis at inference time. $X_1, \ldots, X_N$ represent images of N input modalities and $Y_1, \ldots, Y_K$ represent images of K output modalities. The f represent encoders, parameterised by their respective θ, which map inputs into latent representations. These are aligned with a Spatial Transformer Network (STN) and fused with an operator α. Finally, decoders g, parameterised by ψ, decode the representation into outputs.

representation fusion, and decoding. As each stage is independent, our approach is modular, i.e. encoders and/or decoders can be added to accommodate additional modalities.

4.1.2 Contributions

In summary, our contributions are:

1. We present a novel modular convolutional deep network for MR image synthesis that improves the quality of images synthesised from a single input modality compared to current leading methods. [1]

2. We show that information from multiple inputs can be combined to further improve synthesis quality.

3. By using a single shared decoder for each output modality and a custom loss function, we are able to learn a modality-invariant latent representation to which all input modalities are mapped. This renders the model robust to missing inputs.

4. We demonstrate that the model can be easily extended to new output modalities through the addition of decoders which can be trained in isolation.

[1] Note that comparisons were performed with the state-of-the-art methods at the time of publication.

5. We propose the use of a spatial transformer module [188] for representation alignment. This allows being robust to data misalignment between subjects, reducing the need for co-registered images across modalities.

6. We improve synthesis errors of pathological images by including information from lesion segmentation masks. In this setting, images with synthetic lesions can be generated on request, by adding the affected region as defined by a segmentation mask.

7. We show that our method works for both skull-stripped and non skull-stripped brain data, with no change required, demonstrating that the latent representation is flexible, and not overly tailored to a specific task.

This chapter is organised as follows. Section 4.2 mentions previous work related to ours. Section 4.3 discusses the requirements of a multi-input fusion method. Section 4.4 describes the model details. Section 4.5 describes experimental setup and datasets used. Finally, results are presented in Section 4.6, and Section 4.7 concludes the chapter.

4.2 Related Work

Because of its broad applicability, there has been significant work on MR image synthesis. Previously, an output modality was synthesised from a single input modality, for example by matching input patches to atlases [73]. One main drawback of these approaches is their inability to robustly exploit multiple input modalities. In addition, patch–based methods can be prohibitively slow at test time. Further, the overhead of having many unimodal models from an application standpoint is significant since all these different models have to be trained and maintained. Certainly, there could be a benefit to learning a single multi-purpose model.

Improvements can be made by incorporating information into the input, in addition to raw pixel intensities [71, 78]. In [78] the Location Sensitive Deep Network (LSDN) is proposed, which improves results by conditioning the synthesis on the position in the volume from which the patch comes. Another approach [71], uses random forests to solve the patch based regression problem, and incorporates both multi-scale information and context description features in order to improve the final synthesis. However, although both approaches can fuse information from multiple sources to produce accurate synthetic results, they are not designed to robustly handle missing input modalities.

Although for segmentation rather than synthesis, Hetero-Modal Image Segmentation (HeMIS), a convolutional neural network model, uses a robust fusion method to address the challenge of missing input data [67]. Similar to our approach, HeMIS learns a multi-input mapping to a latent representation in a way that is benefited by, but not dependent on, each of the input modalities. We discuss this approach in more detail in Section 4.4.6, and use their proposed multi-input fusion method as a benchmark in our experiments.

The model we propose here addresses the challenge of multi-input, multi-output synthesis, and does so in a robust way: outperforming existing approaches, and, when inputs are missing, performing as well as a model trained specifically for that fewer input case. Central to our approach to multimodal data is the embedding of inputs into a latent space.

This has been previously approached for example in Correlational Neural Networks [61], which encourage correlated latent representations of multimodal data. However, their current formulation restricts them to the bi-modal setting, due to the use of explicit correlation calculations.

Here we are interested in fusing any number of modalities, and we do not use the formulation of Correlational Neural Network directly. Instead, as our inputs are already similar, in that they are all images of the same organ, and differ only in intensity patterns, we propose a simple method of training that enforces the same constraints: minimising reconstruction error and the distance between the embeddings in the common space, which indirectly maximises the correlation. Thus, our approach is broadly similar to the statistical regularisation approach in [24], in which cross-modal scene representations are learnt. However, in [24], the regularisation is done by encouraging the latent representation activations for all modalities to follow the same distribution. Whereas here, as the various inputs are sufficiently similar, we directly encourage the activations to be equal. Our approach to representation learning is detailed in Section 4.4.5.

4.3 Fusion Requirements

Many synthesis approaches learn to synthesise one modality from another. Thus, when N modalities are being considered, there exist $N(N-1)$ possible one input one output synthesis tasks, and a separate model would be required for each one. This approach not only becomes infeasible as N grows, but also does not benefit from other input sources despite the fact they may be available. On the other hand, if the accuracy of a model is improved by leveraging multiple input modalities, but all inputs are required, the applicability is reduced to only those

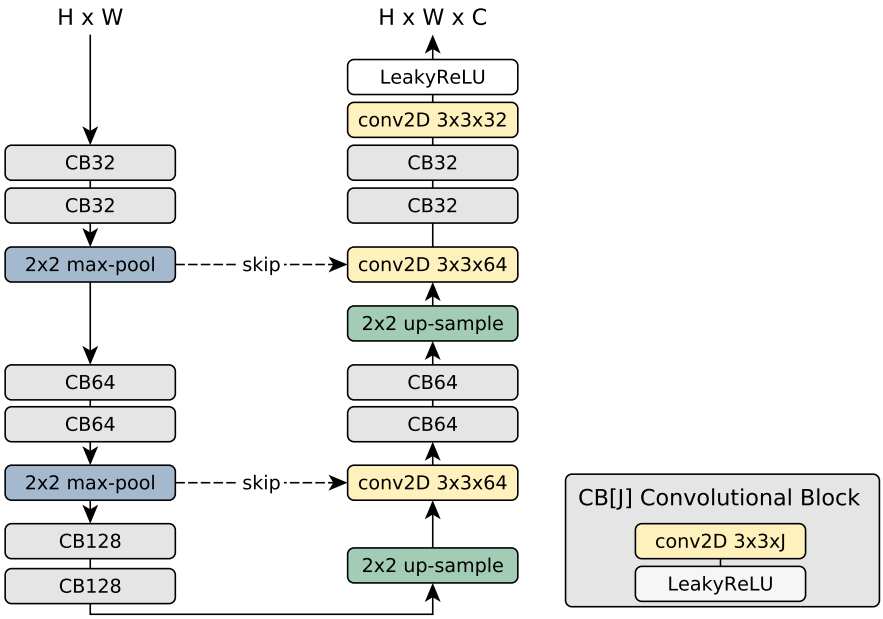


Figure 4.3: Architecture of U-Net [13] like encoder(s) $f(\cdot|\theta)$. Each input modality i has its own encoder, parametrised by θ_i, that maps the input image in modality i to the latent space R_i. We use $C = 16$ channels in the latent space.

situations in which *all* required modalities are available.

The challenge is to build a model which can take as input any subset of the N image modalities to produce its output. We achieve this goal by approaching the task in three stages. Firstly, all inputs are projected into a shared latent representation space, then these latent representations are aligned and fused into a single representation, and finally, mapped to the required output modality. The fusion step, detailed in Sections 4.4.3 and 4.4.6, can be performed on any number of latent representations and having all of the input modalities improves results.

4.4 Proposed Approach

The proposed model is a fully convolutional deep network, that can map multiple input to multiple output modalities. It takes as input 2D volume slices of any subset of modalities, and synthesises the corresponding 2D slices in all output modalities. The model is trained *end-to-end* with gradient descent, and simultaneously learns both encoders and decoders. Through the use of a multi-component cost function the model is encouraged to learn latent representations that balance modality-invariance with the retention of modality specific information. During the

fusion step, the latent representations produced by each of the encoders are combined to form a single latent representation, which is then decoded to produce the final output. Below, we will first describe the four sections of our model in order: encoders, alignment, fusion method, and decoders. We then discuss in depth the importance of learning good representations, and detail a multi-component cost function, providing the motivations for each component.

4.4.1 Encoding

The model learns one independent encoder for each input modality i, with an architecture as shown in Figure 4.3. The encoders embed a single-channel input image $x_i \in X_i$, where $X \subset \mathbb{R}^{H \times W}$ with H, W the image height and width respectively, into a multi-channel latent space. Specifically, the latent representation $r_i \in R_i \subset \mathbb{R}^{H \times W \times C}$ is a C channel image of the same size as the input image. The encoder modules follow a U-Net [13] architecture. The idea behind the U-Net's down-sampling followed by up-sampling and skip connection architecture is to allow the network to exploit information at larger spatial scales than those of the filters, whilst also not losing useful local information. In addition, skip connections facilitate gradient flow during training, as discussed in [200]. An encoder shallower than the original U-Net is used, having only two downsample (and upsample) steps compared to U-Net's four downsample (and upsample) steps. This reduces the training and run times for the model. Although the final quality of synthesis shown herein already outperforms the compared approaches, it may be possible to decrease the error further through the use of deeper encoders. We also replaced the *ReLU* in the standard U-Net with *Leaky ReLU*, as we found that the network is easier to train and it improves the quality of the latent representations.[2] Throughout the network, a stride of 1 is used, and images are padded by repeating the border pixels, so that the final output has the same width and height as the original input. An encoder f is trained for each input modality i to learn the set of parameters θ_i (the network's weights) that fully describes the map from images of the i-th input modality to the latent space R_i. In this model a 16-channel latent representation is used. Experiments with different latent representation sizes showed that this produced good results, whilst keeping the model small enough to easily train (see Section 4.6.1).

[2]One common problem was that the network often got stuck in a bad local optimum when all zero channels in the latent representation developed early in training. The use of LeakyReLUs significantly eased the problem, resulting in consistent performance across runs, likely due to the fact that they always provide a small gradient, whereas ReLUs have 0 gradient when deactivated.

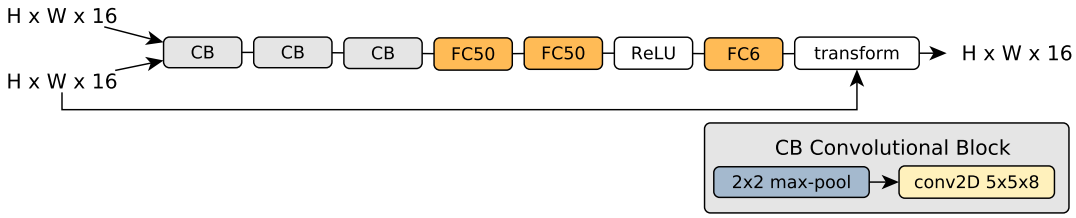


Figure 4.4: The spatial transformer module, that calculates the parameters of an affine transformation used to align latent representations of unregistered images.

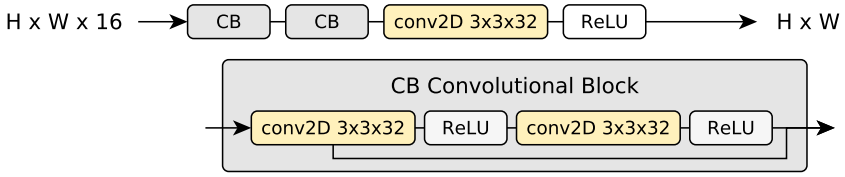


Figure 4.5: The decoder module $g(\cdot|\psi)$, which is built from two residual blocks. Each output modality k has its own decoder, parameterised by ψ_k, that maps latent representations to images of that modality. The channels in the latent space C are set to be 16.

4.4.2 Alignment

The N latent representations can be aligned using the spatial transformer module [188], yielding aligned representations $r_1^*, \ldots, r_N^*$. The spatial transformer stn is a neural network able to learn to apply affine transformations to its inputs, and its architecture can be seen in Figure 4.4. Here, it is used to align all latent representations to the first. To achieve this, the spatial transformer takes r_1 and r_i as input to produce r_i^*, i.e. $r_i^* = stn(r_1, r_i), i \in [2, n]$, where r_i^* is a geometrically transformed r_i. As all other representations are transformed to match r_1, r_1 is left unchanged, and so $r_1^* = r_1$. The parameters of the spatial transformer are learnt implicitly by the overall cost function, see Section 4.4.5.

4.4.3 Fusion

During the fusion step, a fusion operation, α, combines each of the individual representations produced by the encoders into a single fused representation, termed r_α. It is this fusion step that gives the model its robustness to missing input data. In theory, α could be chosen to be any function that takes as input any number of latent representations, and returns a single fused latent representation. This fused representation should integrate information present in the various inputs, such that not only commonly represented features are preserved, but also

unique features expressed in one modality but not the others are retained. Additionally, the fused representation should be robust to varying numbers of inputs and if some input modalities are missing, it should accommodate such missing inputs. Specifically, the aim is that, given any subset of latent representations, a fused latent representation is produced that is at least as good as each of the constituent latent representations, in terms of synthesis quality.

To this end, we use the pixel-wise max function to combine the latent representations into a fused representation. We also consider alternative approaches that include $mean$ feature fusion, HeMIS-like fusion [67], and $mean$ output fusion. These are discussed in detail in Section 4.4.6 and evaluated in Section 4.6.6. All fusion approaches considered are suitable since they do not require a fixed number of inputs. However, fusion approaches other than the max involve all representations, and thus cannot preserve features that are unique to some but not all latent representations. The use of the max means that, in each channel, each pixel of the latent representation has exactly the value of the corresponding pixel in one of the original latent representations. In particular, if the signal is large and positive in one constituent latent representation, then it will be chosen for the fused representation. For N input modalities and corresponding individual latent representations, the fusion operator α is defined as:

$$r_\alpha = \alpha(r_1^*, \ldots, r_N^*) = max(r_1^*, \ldots, r_N^*). \tag{4.1}$$

The fused representation is exactly the same size and shape as each representation r_i. The performance of this fusion method is intimately linked with the nature of the learnt representations, which is detailed in Section 4.4.5. Note that the use of max does not bias the method towards bright final outputs, as the intensities of the synthesised image depend on the decoding step.

4.4.4 Decoding

The decoding stage of the model uses a fully-convolutional network to map the latent representation to a target output modality. Here the input is a multi-channel image-sized latent representation, and the output is a single channel image of the required modality. The exact architecture of the decoder g is shown in Figure 4.5. One decoder is trained for each output modality k, learning the parameters ψ_k, i.e. the network's weights, to map the latent space to the k-th output modality. We kept the decoder shallower than the encoder to encourage the latent representation to contain the useful information in a simple way. Deeper decoders showed no considerable improvement, whilst increasing the computational overhead.

4.4.5 Learning Modality-Invariant Latent Representations

The nature of the latent representation learnt depends critically on the cost function used to train it. The network is trained to minimise a cost function constituted from three cost components, introduced below. The final cost drives the network to achieve three goals:

1. Each modality's individual latent representation should produce all outputs as accurately as possible.
2. The latent representations from all input modalities should be close in the Euclidean sense, and as such be "modality-invariant".
3. The fused latent representation resulting from α should produce all outputs as accurately as possible.

Together these constraints are sufficient to ensure that this architecture works well with a variety of fusion operations, as well as the pixel-wise max approach discussed in 4.4.3.

It is the fusion step that gives the model its robustness to missing input data as the fusion operation α, can be applied to any number of latent representations, and always yields a single fused latent representation. However, the quality of this fused representation depends critically on both the latent representations produced by the encoders, and the nature of this fusion operation. As noted in [61], simply embedding inputs into the same representation space does not ensure that they share a meaningful latent representation. The embeddings, if not encouraged to do so, have no reason to use the latent space in a comparable way. If this is the case, then decoding one latent representation is distinct from decoding the other, and moreover, fusion becomes difficult, as operations such as taking the mean are no longer meaningful. Another way to state this same problem is that, if the different embeddings use the latent space in different ways, then in order to know how to decode a latent representation, you need to know from which modality it originally came, i.e. the meaning of the latent representation is dependent on its initial modality. Thus, in order to overcome this issue we need to produce a latent representation that is *independent* of the originating modality.

Let r_i be the latent representation of image x_i in modality i, i.e. $r_i = f(x_i|\theta_i)$. One requirement of our model is that any input alone should produce good synthesis results, since the model should work well with any subset of inputs, including a single input. Thus, if y_k is the k-th image in a target output modality k, then we want $g(r_i|\psi_k)$ to equal y_k for every input modality

i. Essentially, each modality's individual latent representation should produce all outputs as accurately as possible, when decoded.

Cost component $\mathbf{L_1}$: This desire gives rise to the first cost component. Given N input and K output modalities, the model is fully described by the N encoders, the STN, and the K decoders. We define $\mathbf{L_1}$ as:

$$L_1(f, stn, g) = \mathop{\mathbb{E}}_{x_i, y_k} \left[\frac{1}{K} \sum_{k=1}^{K} \|g(r_i^* | \psi_k) - y_k\|_1 \right], \tag{4.2}$$

where $r_i^* = stn(f(x_1), f(x_i))$ is the aligned representation of image x_i in modality i, and y_k is the corresponding slice in output modality k. Note that we divide by k to average over all outputs. Thus, this cost can be seen as the sum of each modality's average reconstruction error across all outputs. We use the ℓ_1 here instead of the most common ℓ_2 distance, since ℓ_1 helps reduce blurring [11].

Note that decoders, g, are shared, i.e. for each output modality there is exactly one decoder, which is used to decode the latent representations from each of the input modalities. This provides some encouragement for the encoders to come to a shared, modality-invariant representation during training. However, due to the highly non-linear, non-injective nature of the decoder, it is possible for very different latent representations (i.e. ones with a large Euclidean distance between them) to be decoded into very similar output images. Thus, although Equation (4.2) encourages the latent representations to be mutually compatible with a shared decoder, it does not necessarily result in embeddings that *share the same semantics*. In order to ensure that we can meaningfully fuse latent representations, we exploit the fact that the input images are already highly correlated, since they are images of the same subject, and directly encourage the encoders for the different modalities to produce similar embeddings for a given image.

Cost component $\mathbf{L_2}$: To this end, we introduce a second cost that captures the desire that representations from all input modalities should be similar. Although what we really mean by similar here is related to both the details of the fusion operation α and the decoder, we encourage the representations to be close under the Euclidian norm, as if they are sufficiently similar under this metric they will also be sufficiently similar in the required way. In order to bring all latent representations together, we minimise their mean pixel-wise variance (c and p

index the channels and pixels respectively and $r_i^* = stn(f(x_1), f(x_i))$):

$$L_2(f, stn) = \mathop{\mathbb{E}}_{x_i} \left[\frac{1}{|C||P|} \sum_{c \in C} \sum_{p \in P} var(r_1^*[p, c], \ldots, r_N^*[p, c]) \right]. \tag{4.3}$$

Cost component $\mathbf{L_3}$: Although $\mathbf{L_1}, \mathbf{L_2}$ encourage the encoders to learn a shared, modality-independent latent representation, so far there is nothing to encourage this representation to be especially suitable for the fusion operation α used in the model. In fact, so far the particular fusion method chosen has no bearing on the training of the network. The shared representation learnt should be admissible for a wide range of fusion options, but if we decide on a fusion operation in advance, then there is potential to learn a shared representation that works particularly well with that fusion method. As well as meeting the two constraints from above, there may also be sufficient flexibility in the final representation for it to specialise towards the fusion operation in use. To this end, we include a final component in the cost function to directly encourages the minimisation of the reconstruction error from the fused representation:

$$L_3(f, stn, g) = \mathop{\mathbb{E}}_{x_i, y_k} \left[\frac{1}{K} \sum_{k=1}^{K} \| g(\alpha(r_1^*, \ldots, r_N^*) | \psi_k) - y_k \|_1 \right], \tag{4.4}$$

where $r_i^* = stn(f(x_1), f(x_i))$. This is the only cost that involves the fusion operation α.

4.4.6 Other Approaches to Fusion

This multi-component cost function encourages modality-invariant, yet informative, latent representation that can be used with a variety of fusion techniques. Here we discuss alternatives to the pixel-wise max approach (which are also compared with in the experiments).

Latent Mean Fusion: A simple way to fuse latent representations is to average over them. With this approach, the fused representation is the pixel-wise mean of the individual representations:

$$r_\alpha = mean(r_1^*, \ldots, r_N^*). \tag{4.5}$$

This approach should work well if the individual latent representations are approximately noisy versions of a common latent representation. On the other hand, in situations where one input modality can detect details that cannot be seen in the others, this averaging would smooth out these details. Also, it is unable to preferentially select specific input modalities. Therefore, the

information in the latent representation from a highly informative input could be partially lost through averaging with the latent representations from several other less informative inputs.

HeMIS-like Fusion: One approach to the creation of a fused representation, introduced in [67], defines it as the concatenation of the mean and variance of the individual r_i^*:

$$r_\alpha = concat(mean(r_1^*, \ldots, r_N^*), var(r_1^*, \ldots, r_N^*)). \tag{4.6}$$

This method was shown to work very well for image segmentation, producing state-of-the-art results. Our experiments using this fusion showed competitive results also for modality synthesis. HeMIS uses both the mean and variance over the individual representations, and thus the decoder has information about where the representations most disagree, as well as their average value. However, it is still the case that all input representations contribute equally. Unlike max fusion, HeMIS-like fusion can't explicitly rely on more informative inputs. To achieve a 16-channel latent representation with this method we generate eight channels with the encoder, so that the concatenation of the mean and variance is sixteen channels.

Output Mean: As a final baseline, we take the average of the synthesised images decoded from each individual latent representation $r_1^*, \ldots, r_N^*$ independently. Thus, instead of decoding a fused representation to get a single synthesised output, we decode each individual representation into a synthetic image and take the average of those individual images.

4.5 Experimental Setup

This section describes a series of experiments that demonstrate the contributions of the proposed model, and compare with current state-of-the-art methods for medical image synthesis.

4.5.1 Data and Pre-processing

Data from three sources are used in the experimental evaluation: ISLES (Section 2.6.1.1), BRATS (Section 2.6.1.2), and IXI (Section 2.6.1.3). We pre-process data by trimming excess border pixels resulting in volumes of 224×160 pixel images for ISLES, 240×240 for BRATS and 256×256 for IXI. Trimming removes uninformative background areas, and is done in such a way that the resulting image size is divisible by 4, so that the two 2×2 max-pooling, followed by the two 2×2 upsampling operations of the encoder do not change the image size. We keep

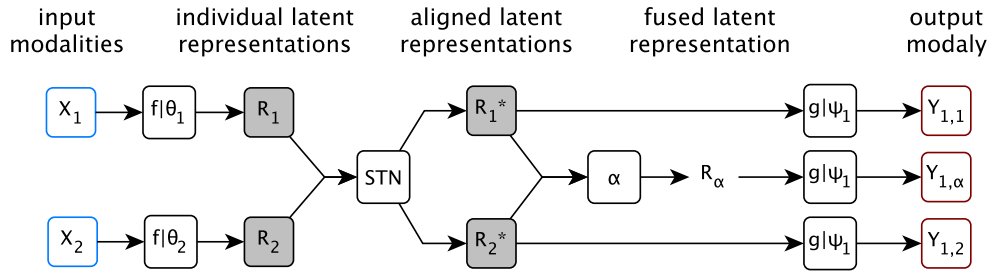


Figure 4.6: The model setup during training for a two input one output case. As we are dealing with a single output there is only one decoder, $g|\psi_1$, used three times: once to decode each of the two individual latent representations R_1, R_2, and once to decode the fused representation R_α. At test time we use the synthesis result from the fused representation as our output. Here we write $Y_{1,i}$ to mean the output synthesised from latent representation R_i.

all slices, which is $\approx$ 150, although the number of slices differs slightly between volumes. As a final pre-processing step we normalise each volume by dividing by the volume's average intensity. As well as centralising all the volumes across all modalities to a mean of 1, this also keeps all values positive, all background values as 0, and maintains the slight differences in volume variance seen between healthy and unhealthy volumes. For the DeepMedic [66] test in Section 4.6.8, we instead normalise the data by subtracting the mean and dividing by the standard deviation, as this is a requirement for the model.

4.5.2 Training and Implementation Details

We train our model w.r.t. a cost function given by the three constituent parts described in Section 4.4.5. The final cost function is:

$$L(f, stn, g) = L_1(f, stn, g) + L_2(f, stn) + L_3(f, stn, g). \tag{4.7}$$

The model is trained using Adam [201] with default parameters. We use a batch size of 16 images. The code is written in Python with Keras [202] and the implementation is available at `https://github.com/agis85/multimodal_brain_synthesis`. We train all models using 5-fold cross-validation. For each cross-validation split, we divide the datasets into training, validation (used to determine when to stop training to avoid overfitting), and test examples. In each fold different test and validation volumes are used, and the remaining

volumes are used for training. In the case of ISLES, the training, validation and test sets consist of 22, 3 and 3 volumes respectively, with one unhealthy volume in each of the validation and test sets, and the remaining 7 in the training set. For BRATS, the training, validation and test sets consist of 42, 6 and 6 volumes respectively, except when using FLAIR images, when we excluded three volumes from the training set as large portions of those volumes were missing in the FLAIR data. For IXI, we use 22 volumes for training, 3 for validation and 3 for testing.

The model at inference time is shown in Figure 4.2. However, during training additional outputs are required by the cost in Equation 4.7, and thus the network has the layout of Figure 4.6.

4.5.3 Benchmark Methods Details

As well as comparing the results of our model with those produced by the fusion approaches discussed in Section 4.4.6 we also compare with three synthesis methods detailed below:

(a) **MP**: Modality Propagation (MP) is a standard synthesis benchmark [73]. We use our own implementation with parameters taken from the original paper. As it is prohibitively slow to synthesise a volume, and it has been shown that the method is outperformed by LSDN [78], we run MP on the ISLES dataset to show that it performs as expected, that is, with a slightly higher MSE than LSDN. See Table 4.2 for details.

(b) **LDSN**: We implemented the Location Sensitive Deep Network (LSDN) as described in [78]. Specifically, we implemented the larger 400,40 neuron version (referred to as LSDN-2 in the paper) without the shrink-connect optimisation, as this is the variant shown to produce the best results in the paper. We train the model to minimise the MSE using stochastic gradient descent with a batch size of 128.

(c) **REPLICA**: Our final baseline method is Regression Ensembles with Patch Learning for Image Contrast Agreement (REPLICA) [71], a supervised random forest image synthesis approach which uses multi-scale features to achieve accurate synthesis results. As this method is able to handle multi-input situations, we compare it to our model in unimodal and multimodal settings. We implemented REPLICA in Python.

4.5.4 Evaluation Metrics

Performance is evaluated with MSE, SSIM and PSNR (defined in Section 3.6) that are calculated at a volume level. Furthermore, we compare our method to the best baseline method in each experiment using a paired t-test and testing for significance at the 5% level. Significant results are shown in bold in the tables.

4.6 Results and Discussion

Here we present the results of a series of experiments examining our proposed model and comparing it to other approaches. In 4.6.1 we first perform experiments to determine the number of channels to use in our latent representation. In 4.6.2 we show the performance of our model on unimodal synthesis. Subsequently, in 4.6.3 we demonstrate that adding inputs increases performance. We also demonstrate robustness to missing inputs comparing against individual models trained specifically for the inputs present. In 4.6.4 we show the importance of each of the three components of our cost function. Next, in 4.6.5, we proceed to demonstrate that we can train a new decoder for an unseen output without learning a new latent representation. In 4.6.6 we show that our model can be used with other fusion methods. In 4.6.7 we demonstrate that our model also works for non skull-stripped data. In 4.6.8 we show that segmentation masks can be used to further improve our model's results, and that they permit the generation of synthetic lesions. In 4.6.9 we show that our model can synthesise images from views not seen during training, and also demonstrate that our synthetic volumes have off-plane consistency. Finally, in Sections 4.6.10 and 4.6.11, we evaluate the STN effect on correcting artificial input misalignments, and test the model generalisability on synthesising new output modalities respectively. Note that the experiments of the final two sections are performed on a smaller resolution.

4.6.1 Latent Representation Size

We first determine experimentally the best latent representation size. Table 4.1 results show that the 16 channel representation outperforms both the 4 and 8 channel versions statistically significantly in both MSE and PSNR, and also by a small margin in SSIM. Although increasing the number of channels beyond 16 may further improve performance, the 16 channel representation achieves the best results while keeping the network's size manageable, and we thus use it for our model in all experiments. Although we could optimally tune the latent representation

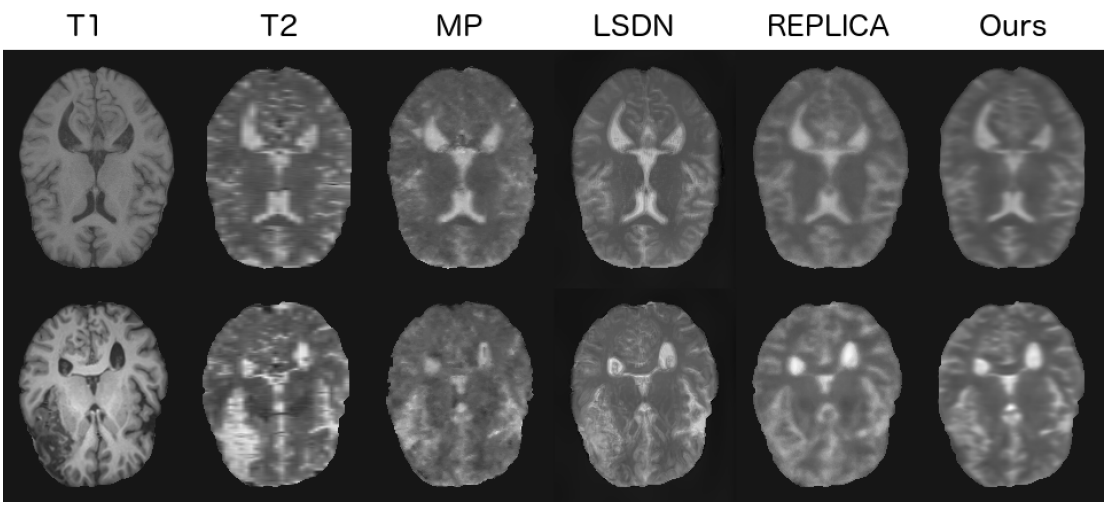


Figure 4.7: Comparison of the unimodal models for $T1 \rightarrow T2$ on a healthy and unhealthy test case. The columns show the input image, the target output image and then the synthesis results of MP, LSDN, REPLICA, and our model respectively. The first row shows a healthy brain, and the second row shows the results on a brain with a large lesion.

	4 channels	8 channels	16 channels
MSE	0.184 (0.07)	0.191 (0.08)	**0.171** (0.06)
SSIM	0.866 (0.02)	0.865 (0.02)	0.869 (0.02)
PSNR	31.61 (1.69)	31.50 (1.72)	**31.10** (1.59)

Table 4.1: Comparison of different sized latent representations for T1, T2, DWI $\rightarrow$ FLAIR.

size for each experimental setup, here we are interested in demonstrating that a single model can perform well in a range of tasks, and thus fix the latent representation size throughout.

4.6.2 Unimodal Synthesis

In our first experiment we train two unimodal models to generate T2 and FLAIR images respectively from T1 inputs. We repeat the experiment for the ISLES and BRATS dataset and compare our models with the benchmark methods described in Section 4.5. The results are presented in Tables 4.2 and 4.3 and show that our model outperforms the other methods. In addition, statistically significant differences are produced on the ISLES dataset for SSIM, and on the BRATS dataset for all metrics. Examples images are shown in Figure 4.7.

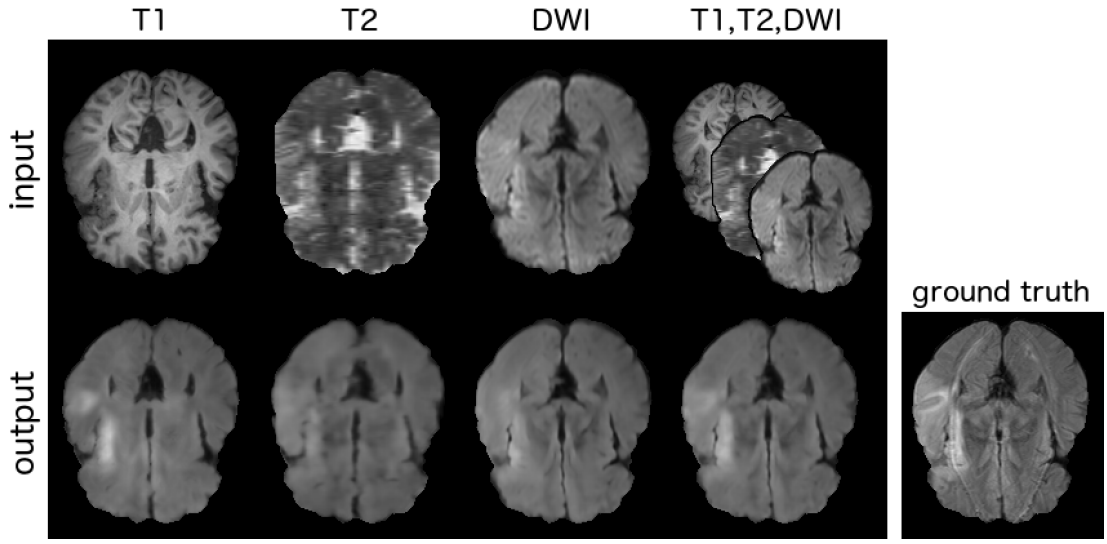


Figure 4.8: Example multimodal synthesis from our model, using all three inputs to synthesise FLAIR. The first row shows the T1, T2 and DWI inputs respectively. In the second row, the images below each input show the synthesis result from that input's latent representation alone (i.e. single input results), the fourth image shows the synthesis result from the fused latent representation, and the final image is the FLAIR ground-truth.

4.6.3 Multimodal Synthesis

To assess the performance of our method on multiple inputs we compare two experimental setups using the ISLES dataset, with T1, T2, DWI as inputs, and FLAIR as output. In *Exper-*

T2	MP [73]	LSDN [78]	REPLICA [71]	Proposed
MSE	0.397 (0.15)	0.345 (0.12)	0.325 (0.12)	0.299 (0.11)
SSIM	0.798 (0.02)	0.811 (0.03)	0.823 (0.24)	**0.831** (0.03)
PSNR	25.22 (0.96)	25.22 (1.36)	25.51 (1.20)	25.78 (1.39)
FLAIR	MP [73]	LSDN [78]	REPLICA [71]	Proposed
MSE	0.343 (0.12)	0.286 (0.10)	0.301 (0.11)	0.268 (0.10)
SSIM	0.802 (0.03)	0.820 (0.03)	0.814 (0.03)	**0.831** (0.04)
PSNR	28.81 (2.13)	29.61 (2.17)	29.43 (2.25)	29.99 (2.24)

Table 4.2: T1 $\rightarrow$ T2 and T1 $\rightarrow$ FLAIR synthesis from unimodal models on ISLES dataset.

T2	LSDN [78]	REPLICA [71]	Proposed
MSE	0.449 (0.12)	0.573 (0.17)	**0.333** (0.13)
SSIM	0.909 (0.02)	0.901 (0.01)	**0.929** (0.17)
PSNR	30.12 (1.62)	28.62 (1.69)	**30.96** (1.85)
FLAIR	LSDN [78]	REPLICA [71]	Proposed
MSE	0.332 (0.16)	0.432 (0.17)	**0.283** (0.14)
SSIM	0.887 (0.01)	0.870 (0.01)	**0.897** (0.01)
PSNR	29.68 (1.56)	28.32 (1.38)	**30.32** (1.61)

Table 4.3: T1 $\rightarrow$ T2 and T1 $\rightarrow$ FLAIR synthesis from unimodal models on BRATS dataset.

iment A we train distinct instances of our model for each possible combination of T1, T2, and DWI inputs, synthesising FLAIR in all the cases. Thus, in total we train 7 different models: 3 unimodal, 3 bi-modal, and 1 tri-modal. As a baseline comparison we also train 7 REPLICA models for the same tasks. In *Experiment B* we take our trained tri-modal model from Experiment A, and at test time, provide different subsets of the inputs (e.g. only T1 images, only T2 and DWI images, etc), to evaluate robustness to missing inputs.

The results of both setups are reported in Table 4.4, and a test example is shown in Figure 4.8. In the table we show in bold results where REPLICA is outperformed with statistical significance. Overall, in all three experiments, we observe the positive effect of multimodal inputs. With our model, this gain does not penalise flexibility as its performance when data is missing (Experiment B) is never worse than the performance of a model trained specifically for the fewer input case (Experiment A). This demonstrates that our model, due to the effectiveness of the latent representation, is able to exploit the input modalities when available, without becoming reliant on them. Our model outperforms REPLICA in 6 of the 7 experimental setups, with statistically significant improvements in 5 cases, when using one model with missing inputs (Table 4.4).

This experiment's setup also allows us to compare our model for different input combinations. Three observations can be made. Firstly, T2 alone gives the highest error, and all other input combinations, (including T1 alone and DWI alone) result in statistically significant improvements over just T2. Secondly, in all two-input cases, the results are better than the results for the constituent modalities individually, and this improvement is also statistically significant in each

Combinations of Input			*MSE (FLAIR modality)*		
T1	*T2*	*DWI*	*REPLICA*	*Proposed: Exp. A*	*Proposed: Exp. B*
✓	—	—	0.301 (0.11)	**0.268** (0.10)	**0.249** (0.09)
—	✓	—	0.374 (0.16)	**0.328** (0.14)	**0.321** (0.12)
—	—	✓	0.278 (0.09)	0.303 (0.13)	0.285 (0.13)
—	✓	✓	0.235 (0.08)	0.215 (0.09)	0.214 (0.09)
✓	—	✓	0.225 (0.08)	**0.208** (0.09)	**0.198** (0.02)
✓	✓	—	0.271 (0.12)	**0.218** (0.08)	**0.214** (0.08)
✓	✓	✓	0.210 (0.08)	**0.171** (0.06)	**0.171** (0.06)
Average:			0.271	**0.244**	**0.236**

Table 4.4: Synthesis of FLAIR images in *Experiment A* and *Experiment B* setups.

case (e.g. when T1 and DWI are given as input the results outperform those for either T1 or DWI alone). Lastly, when T1, T2 and DWI are all provided as input the results are significantly better than in all other cases. To summarise: in all cases adding an additional input modality resulted in a statistically significant improvement, when compared to the results without that additional input. It is worth noting that, as all outputs are coming from the same fixed FLAIR decoder, these significant differences can be understood both as significant differences in the final outputs, and/or as significant differences in the fused latent representations. We also visualise the behaviour of the max-fusion operator α in the three input case, (Figure 4.9). As can be seen, all inputs contribute to the final fused latent representation, and the contributions of the different modalities are not related to tissue classes in a simple way.

4.6.4 Influence of Cost Components

Here we demonstrate that the robustness seen previously stems from the composition of our cost function. To show this, we evaluate the effect of each of the three components described in Section 4.4.5 by assessing the model performance when each component is individually removed. We train three models for synthesising FLAIR from T1, T2, DWI using the ISLES dataset, each with one of the cost components removed. These results, along with the results for training with the full cost function are shown in Table 4.5. The best result is achieved when

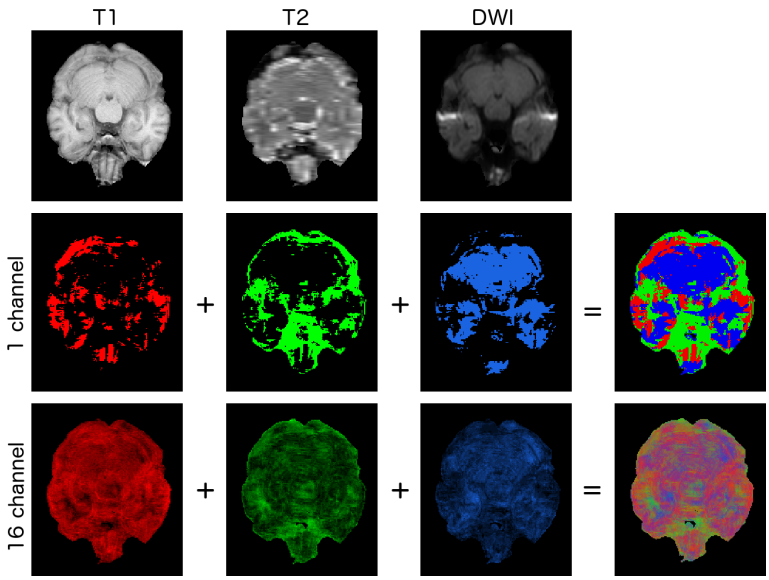


Figure 4.9: Visualisation of the max-fusion behaviour, showing from which inputs the values in the latent representation originate. As can be seen, there is no simple relationship between the input selected and the underlying anatomy. The first row shows T1, T2 and DWI inputs. The first three images in the second row show, for a single channel, the pixels of the individual latent representations that are selected from the max-fusion operator. The fourth image shows the three results simultaneously, with pixels coming from T1, T2 and DWI shown in red, green and blue respectively. The final row is the same as the second row, but rather than showing the results for a single channel, it shows the result averaged over all 16. Note that this figure shows only which inputs are chosen, not the values of the latent representations themselves.

all cost components are employed. Specifically, without L_1 the synthesis result is very good when the model has all inputs, but considerably worse when inputs are missing. Without L_2, the single input results are good, but results with multiple inputs are worse. Finally, when removing L_3, there is a slight degradation in the results with a single missing input, and when all three inputs are given the model is significantly worse. Thus, the multi-component cost, the model achieves high accuracy, whilst retaining robustness to missing data.

The influence of the cost components can also be seen visually in the latent representations learnt by our model, see Figure 4.10. Observe the similarity of all latent representations achieved by minimising their variance through the cost function of Equation (4.3). At the same time the fusion operation α, preserves unique information across the latent components corresponding to bright pixels of the individual latent representations. Note that these bright

Inputs			*MSE (FLAIR)*			
T1	*T2*	*DWI*	*all costs*	*no* L_1	*no* L_2	*no* L_3
✓	—	—	0.249 (0.09)	0.546 (0.19)	0.261 (0.10)	0.250 (0.10)
—	✓	—	0.321 (0.12)	0.903 (0.47)	0.331 (0.14)	0.316 (0.13)
—	—	✓	0.285 (0.13)	0.497 (0.19)	0.293 (0.14)	0.286 (0.13)
—	✓	✓	**0.214** (0.09)	0.324 (0.16)	0.262 (0.12)	0.276 (0.11)
✓	—	✓	**0.198** (0.02)	0.252 (0.10)	0.240 (0.09)	0.228 (0.09)
✓	✓	—	**0.214** (0.08)	0.329 (0.12)	0.345 (0.17)	0.277 (0.10)
✓	✓	✓	0.171 (0.06)	0.185 (0.08)	0.176 (0.07)	0.278 (0.11)
Average:			**0.236**	0.434	0.273	0.273

Table 4.5: Synthesis of FLAIR images when training with different cost functions.

pixels represent strong features, and do not necessarily correspond to bright pixels in the output.

4.6.5 Adding New Decoders

One aim of our latent representations is to introduce modality invariance. This should allow adding inputs and outputs to an already trained network, with minimal performance change. Here we demonstrate that an additional output can be appended to an already trained network. We train a model with inputs T1 and T2, and outputs DWI and FLAIR. At test time, the MSE of DWI images is 0.218. Next, we train another model with the same inputs, but only FLAIR as output; to this already trained model, we add just a DWI decoder that we then train in isolation. The test error for DWI was 0.263, which is $\sim 17\%$ higher, and not a statistically significant difference, compared with the previous case.

4.6.6 Alternative Fusion Operations

In this experiment we demonstrate that our model is still effective with other fusion methods, such as those described in Section 4.4.6. To this end, we train one model for each of these fusion methods with T1, T2, and DWI as inputs, and FLAIR as output on the ISLES dataset. We get the best MSE with our max fusion method, which is equal to 0.171. HeMIS MSE is 0.178, while latent and output mean follow with 0.187 and 0.193 respectively. We also experiment

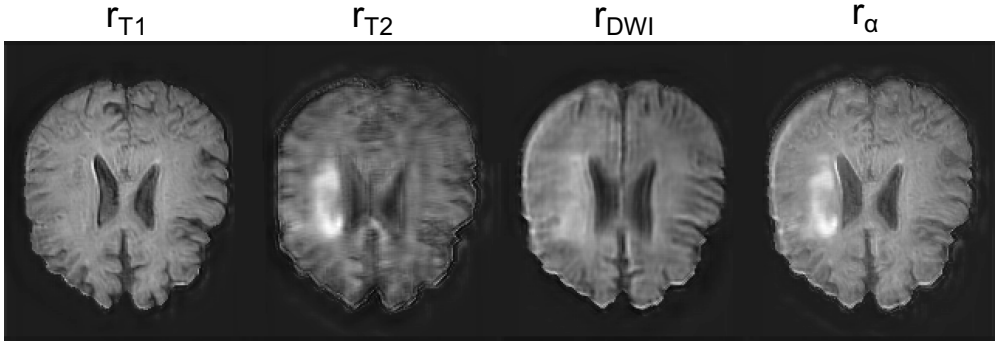


Figure 4.10: A channel from the 16-channel latent representation of our model with T1, T2, DWI inputs. The first three images show the latent representations learnt by the three inputs, T1, T2, DWI respectively. The fourth column shows the fused representation. The high-intensity regions in r_{T2}, which correspond to lesions, are preserved in the fused representation r_{α} despite the latent representations r_{T1} and r_{DWI} showing minimal or no lesion information.

with missing inputs with the HeMIS and latent mean fusion methods. On average, across all seven input combinations, our model achieved an MSE of 0.236 as shown in Table 4.5, whereas HeMIS and latent mean achieved 0.239 and 0.246 respectively, demonstrating that the model still works well with missing inputs in these cases, but performs best with our suggested fusion.

4.6.7 Non Skull-Stripped Data

In these experiments we explore the model in situations where the brain data has not been skull-stripped. As also discussed in [71], synthesising non skull-stripped volumes is difficult because of the intensity inhomogeneity in MR images caused by the dark skull regions surrounded by bright skin and fat regions. REPLICA [71], which is being used as a baseline has been demonstrated to be effective on non skull-stripped data, producing state-of-the-art results, and we compare our method with this approach for evaluation. For this experiment we use 28 volume pairs of PD-weighted and T2 modalities of the IXI dataset. The results are given in Table 4.6. As can be seen, our method outperforms REPLICA, with statistical significance, in all three error metrics. Non skull-stripped example results are shown in Figure 4.11. Although we initially used 28 subjects to be comparable to the ISLES dataset size, to demonstrate that our model scales well and benefits from more training data we trained our model on the full IXI dataset, which consists of 577 volumes (347 training, 115 validation and 115 testing). This significantly improved the performance (compare with Table 4.6), with MSE dropping to 0.067,

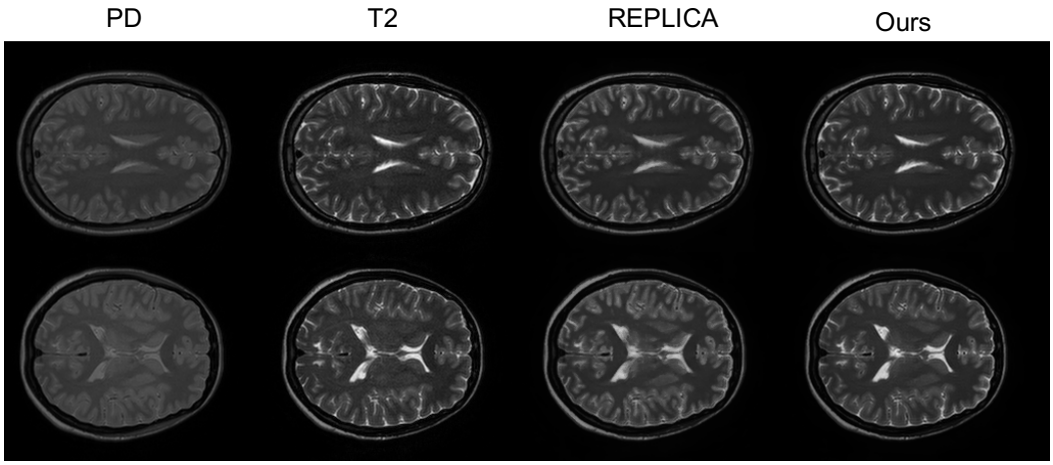


Figure 4.11: Non skull-stripped synthesis examples. The two rows show slices from different test volumes. The columns show the input PD, the ground truth T2, the REPLICA synthetic T2 and our model's synthetic T2 image respectively. Our method produces more accurate outputs.

	REPLICA [71]	Proposed
MSE	0.293 (0.05)	**0.129** (0.04)
SSIM	0.854 (0.03)	**0.865** (0.03)
PSNR	28.93 (1.20)	**32.92** (1.06)

Table 4.6: Results from PD to T2 synthesis on the non skull-stripped IXI dataset.

and SSIM and PSNR rising to 0.872 and 35.20 respectively.

4.6.8 Augmenting Inputs with Segmentation Masks

The ISLES dataset includes segmentation masks that delineate unhealthy regions. We provide the segmentation mask as an additional input channel. With this augmented input, the model can directly modulate its behaviour on affected regions. Specifically, when we train a network with DWI input and FLAIR output, we obtain a MSE of 0.303. When we train a similar network where the mask is provided as an extra channel in the input, the MSE reduces to 0.290. Even though the improvement is $\approx 3\%$, we observed that affected regions in the synthesised images are sharper (also note unhealthy regions are only a small part of a few volumes).

With the same augmented inputs, we can also generate synthetic lesions. To achieve this at

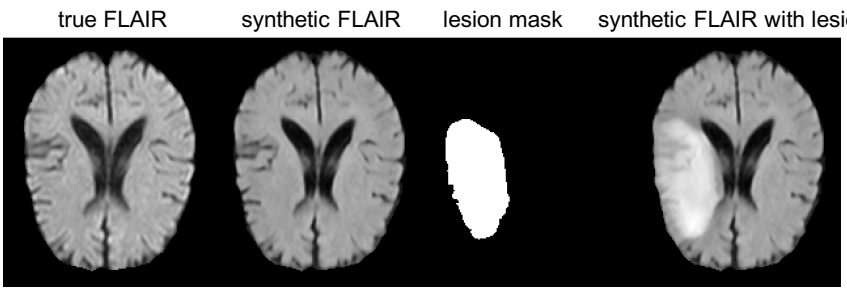


Figure 4.12: Synthesis of a lesion by including a segmentation mask when synthesising an otherwise healthy image. This subject is taken from ISLES dataset in the FLAIR modality.

test time, we use the lesion mask from an unhealthy brain on a healthy brain, and then run the synthesis as normal. A visual example is shown in Figure 4.12. We then train DeepMedic [66] to segment lesions using the FLAIR modality of the ISLES dataset as input. In order to test the quality of our synthetic images, we use DeepMedic to segment the synthetic lesion and get $\approx 84\%$ accuracy (Dice coefficient) on a single test-case.

4.6.9 View-transfer Synthesis

We demonstrate that our architecture can synthesise images (at test time) taken from a different perspective of the 3D volume. Here, we train a model with T1, T2 and DWI inputs and FLAIR output on axial-plane slices as normal, but we test on coronal view slices. An example result is shown in Figure 4.13. Observe that the synthetic image contains all the details including the ischemic lesion, seen in the other modalities and in the ground-truth FLAIR image, visually demonstrating transfer learning capabilities w.r.t. the point of views (axial-coronal planes in this example). Finally, as our method synthesises volumes slice by slice, we evaluate intensity consistency between slices in off-plane reconstructions. As the examples in Figure 4.14 show, consistency is good.

4.6.10 Robustness to Data Misalignment

To examine the performance of our model on unaligned data we trained and tested a model for synthesising FLAIR from T1 and DWI on data in which each T1 volume was randomly rotated about all axes by a number of degrees sampled uniformly at random from $[-8, 8]$ and was shifted randomly on each axis by a (not necessarily integer) number of pixels from $[-2, 2]$. This produced data with misalignment between modalities of the sort that remains after a sim-

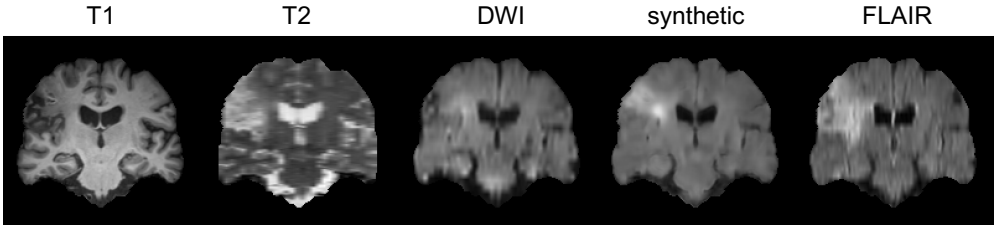


Figure 4.13: A visual demonstration of our model's robustness of to view transfer. We take the model trained on axial-plane slices and test using coronal-plane slices (shown). The image shows the T1, T2 and DWI input slices, the synthesised FLAIR slice, and the ground-truth FLAIR image respectively.

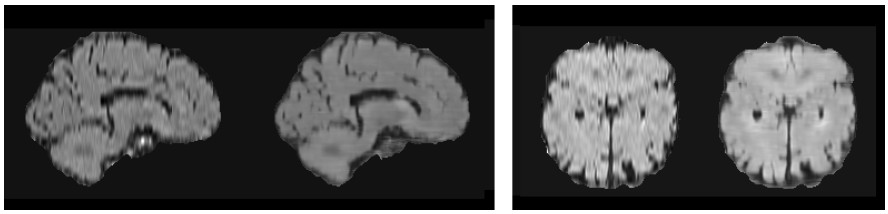


Figure 4.14: Off-plane reconstruction examples. The volume was constructed by synthesising axial slices. Sagittal and coronal slices are taken from this reconstructed volume and compared them to ground truth images. From left to right, the images show a target T1 image, and the off-plane reconstruction, a target FLAIR image, and the corresponding off-plane reconstruction.

ple alignment procedure had been performed. When trained on aligned data, our model and REPLICA achieve MSE of 0.661 and 0.712 respectively, which increases to 0.793 and 0.885 respectively on the unaligned task. However, compared to the unimodal case where only DWI is given as input, which achieves MSEs of 0.821 and 0.901, we observe an improvement of 6% and 2% for our model and REPLICA respectively.[3] Although seemingly a small improvement, rotating and shifting across the z-axis changes the anatomy in the image, necessarily resulting in performance degradation and loss of information by blurring during rotation. However, the model still captures the limited information of the distorted T1 input to improve on the unimodal result.

[3]We remind here that these results correspond to images of lower resolution. The equivalent results for aligned data of full resolution are presented above in Table 4.4.

4.6.11 Transfer Learning

Here we examine the model's ability to generalise to MRI data with different intensity characteristics, not seen during training. We use a model synthesising T2 from T1 trained on BRATS, and test it on ISLES volumes. We first use the model as-is without any fine tuning and get a MSE of 3.990 which we use as a baseline. Then, based on the assumption that the deeper layers of the network are task specific [203], we fine-tune just the decoder using 1, 2 and 3 volumes and get a MSE of 1.439, 1.356 and 1.227 respectively. [4] In addition, fine tuning the decoder is extremely fast, taking ~ 4 minutes on one Titan X GPU.

4.7 Conclusion

In this chapter, we proposed a multi-input, multi-output end-to-end deep convolutional network for synthesis of MR images, capable of fusing information contained in different modalities. Previous synthesis approaches were single-input single-output and thus did not take advantage of the correlated information available within clinical exams. We designed a modular architecture composed of three parts: encoder, latent representation fusion, and decoder. These modules are learnt end-to-end, using a cost function that encourages representations to be modality-invariant, whilst the individual reconstruction error is kept low.

When trained with a single input, our method outperforms the current best methods in all three metrics in each experiment. In particular, significantly outperforming in SSIM in all experiments, and in all metrics on the BRATS dataset. We also demonstrate improved performance on non skull-stripped brain images compared to previous methods. When more inputs are added, the error is further reduced, and our approach is shown to outperform REPLICA statistically significantly in all multi-input experiments. We also show in our experiments that our architecture and cost function can be used in conjunction with various fusion methods, including the one proposed in HeMIS [67]. We also demonstrate that the model is robust to missing inputs: for any subset of inputs it performs as well as a model trained specifically for the subset. Central to our design is the quest towards modality-invariant latent representations. This is achieved via a cost function that aims to unearth shared information whilst still preserving unique (to a specific input) semantics. Such modality invariance has many benefits such as the ability to train new decoders (as demonstrated in 4.6.5).

[4] Note that these results correspond to images of lower resolution.

In addition, the model is also robust to misaligned inputs. In particular, it benefits from multiple inputs, even when they are not well aligned, such that it still outperforms the single input case, even though misalignment means that the slice may well contain different anatomy. Finally, we demonstrated that fine-tuning only the network's decoder on a very small number of volumes allows synthesising volumes from an otherwise unseen data source with high accuracy.

We used MSE, SSIM, and PSNR as evaluation criteria, but these may not directly reflect diagnostic quality. Investigations of new, useful for synthesis, metrics, is an ongoing process in the community. Application-specific metrics are also sought-after and our application driven DeepMedic-based evaluation of pseudo-lesion synthesis points to that direction. This work used three datasets independently, but there is potential for combining information across many sources. This has benefited deep learning in many domains: its application in our context requires suitable pre-processing schemes to alleviate intensity distribution differences between the different sources. Finally, we opted for encoders/decoders that were "small" and fast but still performed exceptionally well. Fine-tuning their design could improve performance further.

Although our approach outperforms the baseline methods in all three metrics, the images produced by LSDN appear sharper than those produced by our method. We believe this is a result of LSDN independently processing small $3 \times 3 \times 3$ voxel cubes to predict a single output voxel. However, although the LSDN approach promotes sharpness, the numerical results show sharpness does not necessarily translate to accuracy: it is certainly possible to have a very sharp, but inaccurate synthetic output.

In summary, this chapter presented a multi-input, multi-output end-to-end deep convolutional network for synthesis of MR images, that was tested on three different brain datasets. We showed that the model is robust, performs well and can handle a variety of different challenges such as robustness to missing input, learning just a new decoder for an unseen modality and even synthesising new (unseen) views of the data. We see that such multimodal models could be well placed to impute data on large databases (e.g. biobanks) w.r.t unimodal approaches. From a deployment perspective they are less complex (one vs many different models to deploy/maintain), more flexible (new outputs can be added with minimal training) and more importantly are robust by taking advantage of information across input modalities, without being reliant on any of them.

Furthermore, we showed that using images as latent variables is suitable for producing multi-

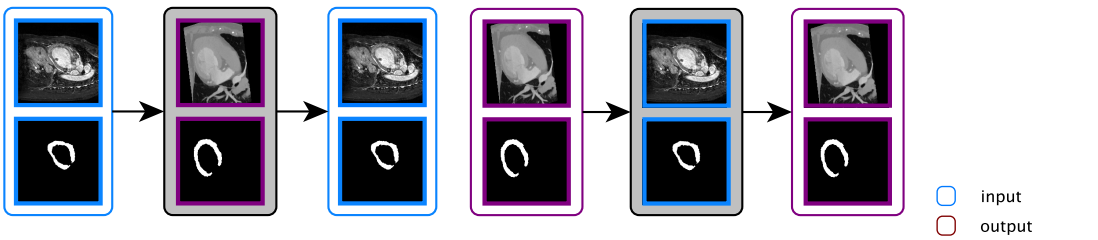
modal representations, as well as for fusing spatial information. Nevertheless, training requires paired data, i.e. pairs of input and output images of the same subject in different modalities. As mentioned in Section 1.1, acquiring such data is challenging in cardiac imaging, and thus prohibits adopting fully supervised approaches for cardiac synthesis. In Chapter 5 we extend this work for unpaired images, using multimodal datasets acquired at different hospitals and containing different subjects. Although, information fusion is not feasible, we will investigate the use of such multimodal data to learn cross-modal correlations for synthesis.

Chapter 5
Cross-Modal Cardiac Synthesis

5.1 Introduction

This chapter extends the work of Chapter 4 on multimodal datasets that contain different subjects, i.e. are unpaired. Specifically, we consider cardiac datasets of MR and CT modalities that are acquired in different hospitals and contain different subjects. In this case we cannot learn a multimodal synthesis model with supervised costs, neither can we learn multimodal representations with similarity costs and fusion as in Chapter 4, since there are no spatial correspondences between the images. Instead, we learn cross-modal synthesis directly from one modality to the other. As seen in Figure 5.1, cross-modal cardiac synthesis is learned through a cycle that maps images and corresponding segmentation masks from MR to CT and back to MR (Figure 5.1a), and vice versa (Figure 5.1b). In this method the role of image representations is taken by a pair of synthetic images and segmentation masks.

As discussed in Chapter 3, techniques for generating synthetic images have undergone significant improvement with the development of GANs. In this chapter, we use CycleGAN to transform unpaired images of one modality into the same image, but in a different modality.



(a) cycle 1: MR to CT to MR synthesis **(b)** cycle 2: CT to MR to CT synthesis

Figure 5.1: Two cycles of cardiac image synthesis between MR and CT modalities.

This chapter is based on:

- Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017, September. Adversarial image synthesis for unpaired multi-modal cardiac data. In International workshop on simulation and synthesis in medical imaging (pp. 3-13). Springer, Cham.

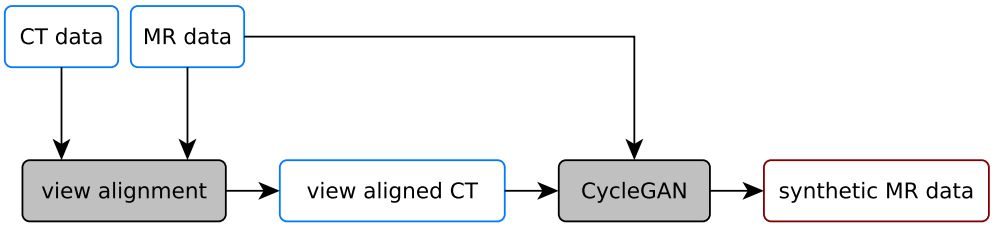


Figure 5.2: A high-level schematic of the synthesis pipeline for cardiac data. The CycleGAN also produces synthetic CT images but here we only use the synthetic MR.

This is particularly useful in cardiac image synthesis, where paired and perfectly aligned images are rare. However, although style transfer for artistic purposes (the original application of CycleGANs) only requires that the resulting images are realistic and maintain semantic content, medical image synthesis is more stringent, and requires preserving precise pixel-level correspondences. We therefore propose a pipeline for directly transforming labelled data into the modality of interest, that incorporates the available labels (segmentation masks) in the translation, such that correspondences between image and labels are preserved in both domains. We demonstrate that the synthetic data consists of examples with potentially beneficial anatomical information. When combined with the original data, this larger and more diverse dataset can then be used to train an improved model for a particular task. Here, we demonstrate this for myocardial segmentation.

5.1.1 Approach Overview

Given two datasets of MR and CT images, the pipeline for our approach is as follows: firstly, we perform a view alignment step, transforming the scale, position and viewing angle of CT images, so that they are broadly the same to the MR images (Section 5.3.1). Secondly, we train a CycleGAN model with adversarial and cycle consistency losses, described in Section 5.3.2, that also includes segmentation masks in training (Section 5.3.3). Once trained, we use the learnt transformation to convert all CT to synthetic MR. A schematic overview of our approach is given in Figure 5.2.

Directly quantitatively assessing the quality of synthetic data when no ground truth exists is challenging. We demonstrate the synthetic data's utility by showing it significantly improves results in a segmentation task.

5.1.2 Contributions

This chapter makes the following contributions:

1. We theoretically explore CycleGAN synthesis in the medical domain and introduce a flexible pipeline for transforming labelled data in auxiliary modalities into labelled data in the modality of interest.
2. We demonstrate that augmenting real with synthetic data significantly improves performance in a segmentation task.
3. We compare our synthetic augmentation with standard augmentation, showing the synthesis approach to be favourable.
4. Finally, we demonstrate a recommended approach, which combines both synthesis and augmentation, and results in the best performance overall.

This chapter is organised as follows. Section 5.2 mentions related work on image synthesis. Then Section 5.3 discusses limitations and presents our approach to cardiac synthesis using a CycleGAN model. Sections 5.4 and 5.5 describe the experimental setup and results, respectively, and finally Section 5.6 concludes the chapter.

5.2 Related Work

To date, there has been very little work on cardiac image synthesis. Our work is based on learning an image transformation function to transfer anatomical information from a source to a target modality. Similar methods have been proposed for cross-modal synthesis of brain images (see Chapter 4), although they require paired and co-registered multimodal datasets. Here, we focus on unsupervised learning of image transformations with no ground-truth target images, which has been revolutionalised by the adversarial training of neural networks [43,48]. Adversarial learning was used for image style transformation in [59], and this method is directly applicable to cardiac data, where there is a lack of paired data.

Although synthesis offers a flexible approach that can be directly applied to expand available data, it is still important to weigh synthesis up, critically, against other approaches. As there is no direct way to measure accuracy when ground truth images do not exist, the value of synthesis

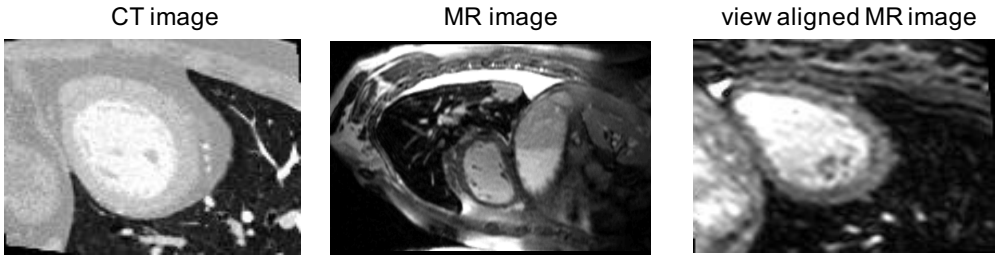


Figure 5.3: An example from the view alignment procedure. The first two images show an original CT and MR slice respectively, and the third image on the right shows the corresponding slice from the view-aligned MR data. Note that, although not co-registered, the first and last images are structurally similar, and essentially differ only in the statistics of the intensities.

should be measured by considering how well it achieves auxiliary tasks. However, this means that synthesis should also be compared with alternative methods for achieving these same goals.

In this chapter we demonstrate the utility of synthesis for improving segmentation via enlarging the set of available training data. Besides synthesis, a dataset can also be expanded using simple geometric augmentation, for example by rotating and reflecting the images. Although simple transformation based augmentation is commonly used to improve results on cardiac segmentation [153, 156], this approach produces derivative examples, and does not benefit from the existence of auxiliary data, which could potentially provide additional real anatomical examples. We directly compare this standard data augmentation with our synthesis approach in Section 5.4, and, as the approaches are not mutually exclusive, we also explore combining both.

5.3 Proposed Approach

We now give step-by-step details of our method, describing the view alignment, the training of the CycleGAN and the generation of the synthetic data. We describe the process in the cardiac setting, using the dataset described in Section 5.4.2, which consists of 40 MR and CT volumes that have been segmented into 7 tissues.

5.3.1 View Alignment

In the view alignment step we make the CT and MR image sets broadly similar in terms of structure. Specifically, we aim to make the layout of the images (the position and size of

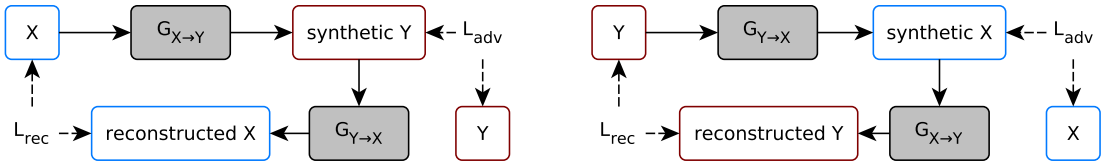


Figure 5.4: The CycleGAN during training. Although both generators occur twice in the graph there is only a single instance of each, which is used in two places. The discriminator costs and reconstruction costs correspond to L_{adv} and L_{rec} respectively as described in Section 5.3.2.

the anatomy for example) not informative as to the dataset from which the image originated. Preventing this is important in order to ensure the adversarial training is effective, otherwise the discriminator may learn to differentiate between real and synthetic data by attending to structural differences, rather than intensity statistics. However, the alignment only needs to be approximate, and any simple registration approach should suffice.

To achieve this we perform an affine transformation on each CT volume to approximately align it to an arbitrary MR volume, re-sampling with tri-linear interpolation to produce an aligned CT volume of the same size as the MR volume (see Figure 5.3 for an example alignment). The exact method we use to align the data is as follows: for each volume we take the labels volume and calculate the centre of mass for each of the 7 classes. This results in a list of 7 3D points for each volume, with each point representing the centre of a particular anatomical region. In order to align the volumes, we calculate the affine transformation that minimizes the squared distance between the corresponding points in two volumes. We then apply that transform to the first volume, and use tri-linear interpolation to re-sample to a 3D array with the same dimensions as the target volume. Any points in the new CT volume that correspond to points outside of the original CT volume are set to 0. Additionally, any points in the MR volume that correspond to points outside of the original CT volume are also set to 0. This again is performed to make the volumes structurally similar, to aid the adversarial training.

5.3.2 Standard CycleGAN and Limitations

Since images are not paired, learning to transform from MR to CT is not straightforward. However, a recent adversarial approach to this difficult task is the CycleGAN: an adversarially trained deep network which simultaneously learns transformations between two datasets containing the same information, but differently represented. It is powerful since it does not

require paired training data, but instead learns via both a discriminator and a cycle loss.

A CycleGAN consists of four networks: two generators $G_{X\to Y}$ and $G_{Y\to X}$, and two discriminators D_X and D_Y. Given two sets of unpaired images, X and Y, the CycleGAN is trained as follows (see Figure 5.4): The generator $G_{X\to Y}$ first transforms images from domain X to Y. The synthetic images are then transformed back to domain X by $G_{Y\to X}$ to complete the cycle. A symmetric cycle in the opposite direction also exists. The training process involves four losses. Two cycle losses, which are direct reconstruction losses between an input image of domain X and the reconstruction produced after completing a cycle. The synthetic images are also encouraged to look realistic by two adversarial losses imposed by the discriminators. More formally, the CycleGAN loss function is defined in Equation 5.1:

$$\begin{aligned} L =& L_{adv}(G_{X\to Y}, D_Y) + L_{adv}(G_{Y\to X}, D_X) + \\ & \lambda L_{rec}(G_{X\to Y}, G_{Y\to X}, X) + \lambda L_{rec}(G_{Y\to X}, G_{X\to Y}, Y), \end{aligned} \tag{5.1}$$

where λ is a hyperparameter set to $\lambda = 10$, as in the original paper [59]. Given input and output samples $x \in X$ and $y \in Y$, respectively, the first cycle and adversarial losses are defined in Equations 5.2, and 5.3:

$$L_{rec}(G_{X\to Y}, G_{Y\to X}) = \mathop{\mathbb{E}}_{x,y} \left[\| G_{Y\to X}(G_{X\to Y}(x)) - x \|_1 \right], \tag{5.2}$$

$$L_{adv}(G_{X\to Y}, D_Y) = \mathop{\mathbb{E}}_{x,y} \left[D_Y(G_{X\to Y}(x))^2 + (D_Y(y) - 1)^2 \right]. \tag{5.3}$$

The losses for the second cycle with $y \in Y$ and $x \in X$ being the inputs and outputs are similarly defined. Here the adversarial loss corresponds to the Least-Square loss of [50]. The discriminators are trained by maximising L_{adv}.

We apply CycleGAN to learn to transform a CT image into a synthetic MR image that cannot be recognised as synthetic by a discriminator network. At the same time, the synthetic MR image must be able to be accurately converted back into a CT image, as similar as possible to the original CT image, via another learnt transformation. Thus, the synthetic MR image, whilst appearing realistic, must also retain relevant information from the CT. This encourages the synthetic MR to contain the same anatomy as is present in the input CT.

CycleGAN Limitations for Medical Applications

Initially, we applied the CycleGAN directly to the MR and CT images. However, we found that although the resulting images were promising in terms of realism, the myocardium in the synthetic image was frequently shifted and deformed during a modality transformation (see Figure 5.5). As a result, the synthetic MR data had no accurate labels, as we could not assume the label was the same as in the input image. Two properties of CycleGAN cause this effect.

Deterministic transformations: By design CycleGAN's generators learn deterministic transformations, i.e. the same input image will always yield the same output image. Thus, in cases where information is present in one modality, but not in the second, it must be *deterministically invented* by the transformation. For example, if in MR lungs do not have strong signal (air has poor contrast), then the network has to *realistically invent* plausible signal for the CT. Conversely, from CT to MR, the network will have to remove this signal but due to cycle-consistency it has to then somehow add it back. So either the network weights must encode this transformation or somehow the image synthesised must contain this information.

Fixed and altered image properties: A transformation between images will change some properties of the input image, and leave others unchanged. CycleGAN implicitly captures this split between properties. There is *no* explicit delineation of the two property types, instead the transformed image must be indistinguishable from a real image. Thus, even in the CycleGAN's theoretical best-case when the distribution of synthetic images is identical to the distribution of real images, the properties that change and the properties that are fixed by the transform are not deducible. In other words, even knowing that the CycleGAN is working as well as possible, it is still not possible to infer what exactly are the transformations it is doing.

Although for some applications this is acceptable, in medical image analysis understanding the precise operation of the transformations is key. In our experiments small shifts occurred with reasonable frequency. It appears that even the authors of [59] alluded to some of these issues in their manuscript and project website, discussing tasks that require geometric transformations. The effects of the issues became apparent when CycleGAN-driven synthesis was used for the first time in quantitative tasks since even such small shifts can cause problems in tasks of segmentation and registration.

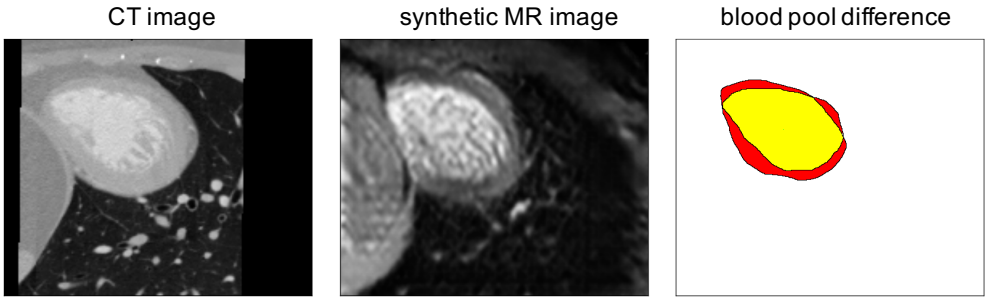


Figure 5.5: Example of spatial misalignment between an input CT image (left) and corresponding synthetic MR image (centre). Although the synthetic MR image has realistic intensities, the alignment with the original CT image is only approximate. The right most image shows the difference between the inner contours of the blood pools of the two images. The area where they agree is shown in yellow, and where they disagree is shown in red.

5.3.3 CycleGAN with Masks

For synthesis to be maximally useful, the anatomy in the synthesised images must be perfectly aligned with the anatomy in the input images. However, the standard CycleGAN can warp the anatomy during the transformation even when producing realistic images (see Figure 5.5). This can be prominent when using small datasets for training, such as in our case (see Section 5.4.2), in which the image variability is small. Due to the large capacity of the generator networks, anatomy shifting could also be attributed to the network memorising images in the source and target modalities.

To mitigate this issue, a need to regularise the geometric transformations emerges. Other approaches [167], trained a segmentation algorithm for one modality in unison with the CycleGAN, as an additional supervised task. On the contrary, we included both the mask of the myocardium and the image as two channel inputs to the CycleGAN, such that it learnt to transform CT images and their corresponding myocardium segmentation mask into realistic MR images and corresponding segmentation masks. This did not stop the anatomy shifting during the transformation, but meant that we still had accurate (synthetic) labels for the synthetic images. A schematic of this approach can be seen in Figure 5.6.

We apply the mapping learnt with the CycleGAN to the view-aligned CT images and masks, producing a synthetic MR image and mask for every CT sample in the dataset. The result is a synthetic labelled dataset of MR cardiac images, which can be used for any task of interest.

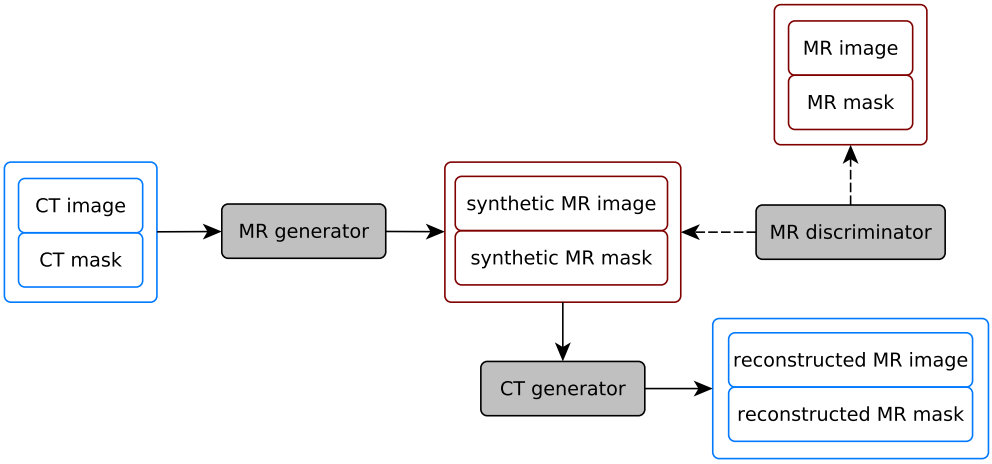


Figure 5.6: Unfolded CycleGAN training for CT to MR synthesis: a CT image with its segmentation mask is mapped to a synthetic MR image and mask by a generator network. A MR discriminator then tries to discriminate real from synthetic MR. The CT and Mask are also reconstructed form the synthetic MR by a second generator network, which aims to reconstruct the original CT exactly. The generator learns both by trying to fool the discriminator, and by minimising the discrepancy between the real CT and its reconstruction.

5.4 Experimental Setup

In this section we examine the effect of synthetic results in the accuracy of myocardium segmentation. We train a segmentation model, detailed in Section 5.4.1, on various combinations of synthetic and real data, with and without augmentation and report the Dice coefficient, described Section 3.6, on 3-fold cross validation. The data and pre-processing steps are described in Section 5.4.2, and the experimental details in Section 5.4.3.

5.4.1 Segmentation

To segment the images, we train a neural network with an architecture similar to the U-Net [13]. Specifically, the network consists of 3 downsample and 3 upsample blocks with skip connections between each block of equal size filters. This architecture was chosen as similar fully convolutional networks have been shown to achieve state-of-the-art results in various segmentation tasks, including cardiac, and U-Net is a standard benchmark approach. Here we have not specifically optimised the architecture or hyperparameters for the segmentation task being considered, since the aim is to evaluate the synthetic results. Our model is implemented in

Keras [202] and trained using Adam [201] with batch-size 16 and an early stopping criterion, based on the validation data, to avoid overfitting.

5.4.2 Data and Pre-processing

The experiments use data from MM-WHS that is described in Section 2.6.2.1. We centered the anatomy (the bounding box of the labelled anatomical regions) within the MR volumes, and trimmed each volume to 232×232, padding with 0s where necessary, but maintaining the native resolution. Then, for each volume, we clipped the top 1% of pixel values and re-scaled the values to $[-1, 1]$. Finally, we removed slices that did not contain myocardium, resulting in 20 volumes with an average of 41 slices per volume (816 slices in total). For the cardiac CT data no centering or trimming was necessary, as the data is aligned with the MR data in the view alignment step of Section 5.3.1. However, we again clipped the top 1% of values, and scaled the values to $[-1, 1]$.

5.4.3 Experiment Details

Below we detail the five experiments we used to evaluate the quality of the synthesised cardiac MR data. We repeated all experiments on three different splits of the data, each time training a CycleGAN on 15 MR and 15 CT volumes, and then training the segmentation network described in Section 5.4.1. In every split, the 5 MR volumes used for testing the segmentation network were excluded, as were the 5 CT volumes which were aligned with them in the view alignment step. Thus the final test volumes have not been used anywhere in the pipeline. Out of the remaining 15 MR volumes, we used 10 for training and 5 for validation.

(a) **Real:** Firstly, as a baseline we train the segmentation network on 10 real MR volumes, using the other 5 MR volumes for validation, and obtain a mean test Dice of 0.613.

(b) **Synthetic:** Secondly, to directly evaluate the quality of synthetic data, we train the segmentation network on 10 synthetic volumes, validating on 5 synthetic volumes. We then test the final model on the 5 real MR volumes and obtain a Dice coefficient of 0.580.

(c) **Real and Synthetic:** Next we combine the real and synthetic data and train the segmentation network on a total of 25 volumes (10 real and 15 synthetic), again using 5 real volumes for validation. This combined training gives a performance gain of $\sim 15\%$

training data	split 1	split 2	split 3	average	relative to real
just synthetic	55.3	51.6	67.2	58.0	0.946
just real	58.4	61.3	64.2	61.3	1.000
augmented real	63.2	68.5	71.1	67.6	1.103
real and synthetic	**65.7**	69.9	**75.7**	70.4	1.148
augmented real and synthetic	65.0	**73.8**	74.8	**71.2**	**1.161**

Table 5.1: Dice scores (%) of U-Nets trained on various data combinations. In all cases the model is evaluated on real MR images.

compared to training on real data alone.

(d) **Augmented Real:** Next we augment the real data using horizontal and vertical flips generating a total training set of 25 volumes (10 real 15 flipped) to allow direct comparison with synthetic augmentation.

(e) **Augmented Real and Synthetic:** Synthesis and data augmentation are not mutually exclusive and can be simultaneously used along with the existing real MR data. We therefore combine the real and synthetic training data, and also use horizontal and vertical flips to expand the data to double the size. This results in 50 training volumes, and we again use 5 real volumes for validation during training.

5.5 Results and Discussion

All results are presented side-by-side in Table 5.1. In addition, Figure 5.7 provides examples of synthetic results. The first observation is that using just the synthetic data is almost as good as using the real data, in terms of resulting segmentation, only resulting in a 5% loss of accuracy and this difference is not statistically significant at the 5% level. This is likely the result of small errors present in the synthetic images. Next, it is informative to compare real data with standard augmentations against the combined real and synthetic data. In both cases the segmentation algorithm was trained on 25 volumes, including the same 10 real volumes, and both approaches improve the final segmentation accuracy with synthetic and geometric augmentation leading to 14.8% and 10.3% improvements respectively. Finally, when the real and synthetic data is combined, and geometric augmentations are also applied, the greatest improvement is seen,

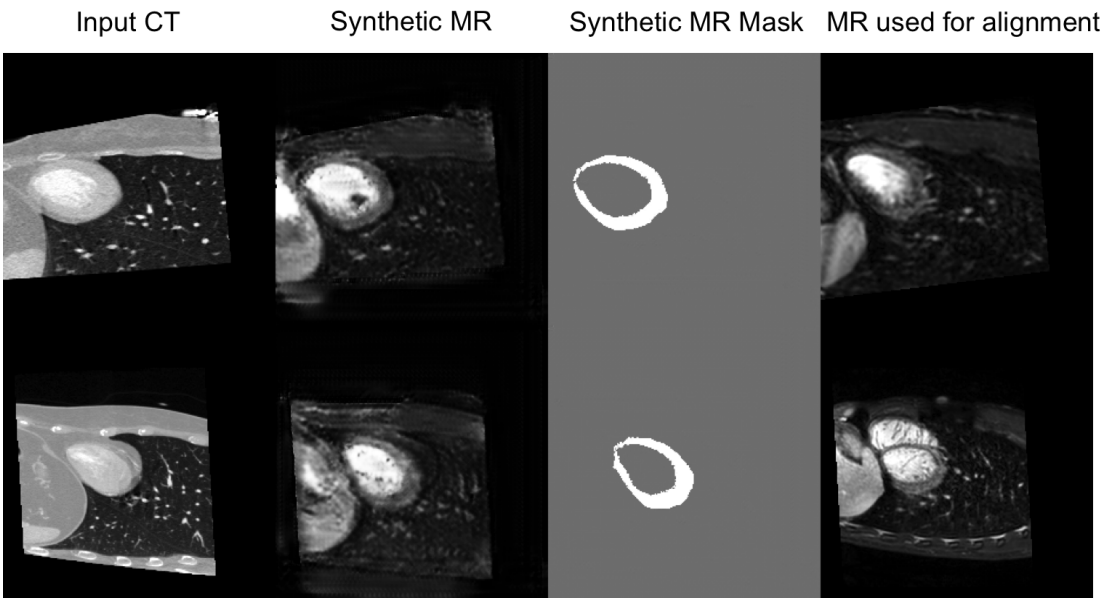


Figure 5.7: Two examples of MR synthesis. From left to right it is shown, the real CT image, the resulting synthetic MR image, the synthetic segmentation mask and finally the real MR image of the volume to which the real CT volume was aligned in the view alignment step. Note that the shape and position of the myocardium is similar but not identical between the CT input and corresponding synthetic MR output. Also, observe that in the upper row the synthetic data contains a dark artifact within the ventricle.

with a 16.1% increase in accuracy over the baseline.

The difference in performance between the real and synthetic data, and just the real data is significant at the 5% level, as is the difference between the real and synthetic data and the augmented real data. Further, adding augmentation to the real and synthetic data does not lead to a statistically significant improvement.

5.6 Conclusion

We have demonstrated that it is possible to produce synthetic cardiac data from unpaired images coming from different individuals. Moreover, these synthetic images are accurate enough to be of significant benefit for further tasks, either used alone or to enlarge existing datasets. Specifically, we have shown that it is possible to produce synthetic cardiac MR images from cardiac CT images, and that these images can be used to improve the accuracy of a segmentation algo-

rithm by 16% when used in combination with standard geometric augmentation techniques. We also demonstrated that the synthetic data alone was sufficient to train a segmentation algorithm only 5% less accurate than the same algorithm trained entirely on real data.

As can be seen in the results, the largest gains are made when the synthetic data is included in the training set, suggesting that new anatomy, containing additional examples of real structure and natural local variations, being introduced from the auxiliary data is most beneficial for improving results.

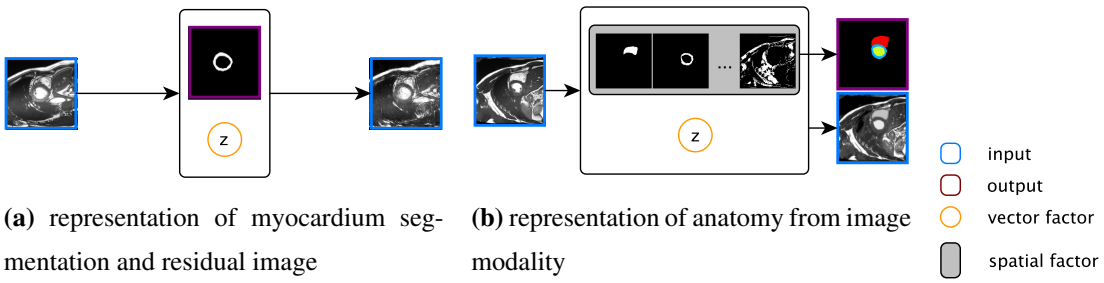
Finally, and as discussed in Section 5.3.2, initial attempts to train the CycleGAN on images alone resulted in synthetic images that were not aligned with the mask of the image from which they were synthesised, meaning that the synthetic images were no longer accurately labelled, and so could not be evaluated through training of segmentation algorithms as above. Here, we overcame the issue though the inclusion of the myocardium mask as input to the CycleGAN, which resulted in accurately labelled synthetic images. However, this unveils the ability of CycleGAN architectures to introduce transformations during translation.

Further problems arise when the information capacity between the translation domains is different, for example when one domain is a categorical segmentation. Chapter 6 demonstrates this limitation, as well as how neural networks attempt to "invent" or "hide" information in order to achieve such translations. In Chapter 6 we also propose disentangled representations as a way of introducing auxiliary variables to overcome this information loss. This renders cyclic (reconstruction) constraints useful in segmentation tasks, and creates the potential for using such representations in other medical image analysis tasks.

Chapter 6
Disentangled Representation Learning

6.1 Introduction

Similar to single-input, single-output synthesis discussed in Chapters 4 and 5, this chapter considers image segmentation as a synthesis task, in which the input domain consists of cardiac images but the target domain is rather a semantic map of the heart. Similar to Chapter 4 we use paired data (supervised) to learn a translation from images to semantic maps (segmentation), but simultaneously, as in Chapter 5, we also use unpaired images, i.e. unlabelled, to improve performance (unsupervised). Since the two domains differ in information capacity, we approach the task from a representation learning view, and propose two approaches in Sections 6.4 and 6.5 with schematics illustrated in Figure 6.1 that respectively learn a representation of the myocardium and residual anatomical information (Figure 6.1a) and a representation of anatomical and imaging information (Figure 6.1b).



(a) representation of myocardium segmentation and residual image

(b) representation of anatomy from image modality

Figure 6.1: Two representations of cardiac images in disentangled spatial and vector factors.

This chapter is based on:

- Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R. and Tsaftaris, S.A., 2019. Disentangled representation learning in cardiac image analysis. Medical Image Analysis, 58, p.101535.
- Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D., Dharmakumar, R., Tsaftaris, S.A., 2018. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 490-498). Springer, Cham.

In representation learning, latent variables must be maximally informative for the task at hand, whilst being invariant to unrelated information (e.g. variations in imaging and noise), so that they can generalise to unseen examples [17]. Invariance to some factors, e.g. translations, can be attributed to the architecture, for instance with the use of convolution and max-pooling, but invariance to more complex factors is achieved by the learning process, and can be encouraged with regularisers. At a high level the aim is to keep relevant but discard irrelevant information, however which information is relevant is strongly task dependent. We therefore consider disentangling representations into meaningful components (factors).

Disentangled representations offer many benefits. For example, they ensure the preservation of information not directly related to the primary task, which would otherwise be discarded, whilst they also facilitate the use of only the relevant aspects of the data as input to later tasks [17]. They also have considerable potential in the analysis of medical data. In this chapter we combine recent developments in disentangled representation learning with strong prior knowledge about medical image data: that they necessarily contain information on the anatomy and the image modality.

6.1.1 Approach Overview

We propose two models for learning different decompositions of anatomy and modality using spatial and non-spatial factors. The first model, Spatial Myocardial Disentanglement Network (SMDNet), decomposes input images into a segmentation map of the myocardium (spatial factor) and a latent vector of image intensity and surrounding anatomical information (non-spatial factor), and is presented in Section 6.4. The second model, Spatial Disentanglement Network (SDNet), is more generic and decomposes input images in a semantic anatomical map (multi-channel spatial factor) and a latent vector of only image intensity information (non-spatial factor), and is presented in Section 6.5.

In both models, part or all anatomical information is represented spatially (as a semantic map) to maintain pixel-level correspondences with the input. As we demonstrate below, a spatial anatomical representation is useful for various modality independent tasks, for example in segmentation (single-class myocardium segmentation in Section 6.4.3.2 and multi-class cardiac segmentation in Section 6.5.3.1), as well as in calculating cardiac functional indices (Section 6.5.3.2). Disentanglement of anatomy and modality factors in SDNet also allows a meaningful representation of the anatomy that can be generalised to any modality, and provides a

suitable format for pooling information from various imaging modalities.

In both models the non-spatial factor contains the spatial factor's residual information, such as global image modality information, specifying how the anatomy is rendered in the final image. Maintaining a representation of the modality characteristics allows, among other things, the ability to use data from different modalities (Section 6.5.3.3). In SMDNet the non-spatial factor further encodes information of the anatomical structures surrounding the myocardium. Encoding this residual information within the non-spatial factor most importantly enables reconstructing the input, which is key in utilising unlabelled data for semi-supervised learning.

Finally, the ability to learn this factorisation using a very limited number of labels is of considerable significance in medical image analysis, as labelling data are tedious and costly. Thus, it will be demonstrated that the proposed factorisations, in addition to being interpretable, lead to considerable performance improvements in (single-class and multi-class) segmentation tasks when using a very limited number of labelled images.

6.1.2 Contributions

In summary, our contributions are the following:

1. We propose new methods for disentangling images into a spatial map and a continuous vector, which are directly applicable to medical images for representing anatomical and non-anatomical information. We also apply constraints on the spatial representation to be semantically meaningful, so that it corresponds to one or multiple anatomical regions.

2. We demonstrate the utility of our methods in a semi-supervised segmentation task and on different datasets, and show that we maintain a good performance even when training with labelled images from only a single subject.

3. We show properties of the decomposed latent space by generating examples using latent space arithmetics.

4. We show that a semantic anatomical representation is useful for other anatomical tasks, such as inferring the Left Ventricular Volume (LVV). More critically, we show that we can also learn from such auxiliary tasks demonstrating the benefits of multi-task learning, whilst also improving the learnt representation.

5. Finally, we demonstrate that disentangling anatomy and modality factors enables multi-modal learning, where a single encoder is used with both MR and CT data, and show that information from additional modalities improves segmentation accuracy.

This chapter is organised as follows. Section 6.2 discusses previous work related to disentangled representations and semi-supervised segmentation. Then, Section 6.3 presents the benchmarks used in the experimental evaluation our methods. Sections 6.4 and 6.5 present the two proposed approaches respectively, and finally, Section 6.6 concludes the chapter.

6.2 Related Work

Here we review previous work on disentangled representation learning, which is typically a focus of research on generative models (Section 6.2.1). We then review its application in domain adaptation, which is achieved by a factorisation of style and content (Section 6.2.2). Finally, we review semi-supervised methods in medical imaging in Section 6.2.3.

6.2.1 Disentangled Representation Learning

Interest in learning independent factors of variation of data distributions is growing. Several variations of VAE [12, 118] and GAN [119] have been proposed to achieve such a factorisation. These methods learn disentangled representations in terms of continuous or discrete variables; however, spatial information could be directly represented in a convolutional map, and this would be useful when the learning task is semantic segmentation. Our proposed methods produces a decomposition as a combination of spatial and non-spatial information. This makes our learned representation directly applicable to segmentation tasks.

6.2.2 Style and Content Disentanglement

Our approach here can be seen as similar to a disentanglement of an image into style and content, where we represent content (i.e. in our case the underlying anatomy) spatially. Concurrent to our approach, there have been recent disentanglement models that also use vector and spatial representations for the style and content respectively [25, 130–132]. The intricacies of medical images differentiate us by necessitating the expression of the spatial content factor as categorical in order to produce a semantically meaningful (interpretable) representation of the anatomy,

which cannot be estimated and rather needs to be learned from the data. This discretisation of the spatial factor also prevents the spatial representation from being associated with a particular medical image modality. The remainder of this chapter uses the terms anatomy and modality to refer to the synonymous content and style.

6.2.3 Semi-supervised Segmentation

A powerful property of disentangled representations is their ability for semi-supervised learning [21]. An important application in medical image analysis is (semi-supervised) segmentation. Semi-supervised segmentation has been proposed for cardiac image analysis using an iterative approach and conditional random fields post-processing [171].

6.3 Benchmark Methods

In Sections 6.4, and 6.5 that demonstrate semi-supervised segmentation, we use the following benchmarks for comparison.

(a) We use **U-Net** [13] as a fully supervised baseline because of its effectiveness in various medical segmentation problems, and also since it is frequently used by participants of cardiac challenges, such as MM-WHS and ACDC. It's architecture follows the one proposed in the original paper.

(b) As a semi-supervised benchmark, shorthanded as **UNetGAN** below, we add an GAN with a mask discriminator to the U-Net's supervised loss, to allow adversarial training [178]. This is useful when there are images with no ground truth masks, although learning to produce a segmentation mask does not guarantee preserving spatial correspondence between the input image and the generated masks.

(c) We also use the **self-train** method of [171], which proposes an iterative method of using unlabelled data to retrain a segmentation network. In the original paper a conditional random field post-processing is applied. Here, we use U-Net as a segmentation network (such that the same architecture is used by all benchmarks) and we do not perform any post-processing for a fair comparison with the other methods we present.

6.4 Spatial Myocardial Disentanglement Network

We propose a Spatial Myocardial Disentanglement Network (SMDNet), that disentangles images into a myocardial segmentation mask, and a latent vector of intensity and residual anatomical information. Specifically, we train two networks: one that learns a decomposition into spatial and non-spatial latent factors, and one that learns to reconstruct the input image using the decomposed representation. We demonstrate our method in semi-supervised myocardium segmentation, using a small amount of labelled but a large pool of unlabelled cardiac cine-MR images. In this application, our method learns to decompose the shape and location of the myocardium from information related to surrounding structures and pixel intensities (related to scanner properties and other imaging characteristics).

6.4.1 Materials and Methods

6.4.1.1 Motivation

A useful latent representation is one that describes the data well. Spatial (segmentation) maps can be considered a form of latent variable that allows visual inspection of what a network learns. At the same time, an easy (unsupervised) way to see whether a latent representation captures the data is to use a decoder to reconstruct the input. In fact, even CycleGANs are autoencoders: they encode (and decode) the input via an intermediate output and thus inspire the design of our approach. Yet they have problems particularly when the intermediate output is discretised (a binary mask) and supervised losses are introduced. Their performance heavily depends on the weighting of the losses, as shown in Figure 6.2. If the segmentation loss is weighted higher than the reconstruction loss, it is not possible to reconstruct the input since the binary mask does not contain enough information for the transformation. When differently weighted, information is stored in the binary mask ruining semantics. This confirms findings of others, that a CycleGAN resolves the many-to-one/one-to-many problem by storing low-frequency information in the output image [204]. We can see that the two losses are antagonistic, and a standard CycleGAN is not suitable as is. We need to introduce variables that break the many-to-one problem, encouraging a balance between the losses to achieve good segmentation and reconstruction.

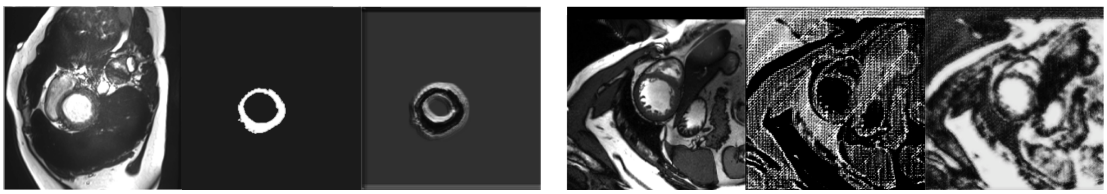


Figure 6.2: Input images, segmentation masks and reconstructions produced by a CycleGAN. Left: high weight on segmentation, right: high weight on reconstruction.

6.4.1.2 SMDNet

Our model can be seen as an encoder-decoder and is comprised of two interconnected neural networks, a "decomposer" and a "reconstructor", as illustrated in Figure 6.3. The former decomposes an input 2D image (slice in a cine acquisition) into two components: a spatial representation of the myocardium in the form of a binary mask, and a latent representation of the remaining anatomical and imaging features in the form of a vector. Thus, the mask is an image having pixel to pixel correspondences with the input and is inherently spatial, whereas the other representation is a vector representing information in a high level way that is not directly spatial. The reconstructor receives the two representations and aims to synthesise the original input image. Given a successful decomposition, the binary mask acts as a guide defining where the reconstructed myocardium should be. The role of the vector component is then to learn some topology around the myocardium and fill the necessary intensity patterns, and allow for many-to-many mappings.

Costs

More formally, let f and g be the decomposer and reconstructor. Given an image slice $x \in X \subset \mathbb{R}^{H \times W}$, where H and W are the image height and width respectively, we aim to learn weights of f to decompose into a mask m and a 16 dimensional vector z, that is $f(x) = \{f_M(x), f_Z(x)\} = \{m, z\}$, and the weights of g to remap the decomposition back to an image $g(f_M(x), f_Z(x))$.

In a semi-supervised setup data comes from a labelled set $\{X_L, M\}$, where $M := \{0, 1\}^{H \times W}$ is a set of segmentation masks for images X, and an unlabelled set X_U where usually $|X_U| > |X|$. We now define the following losses. Firstly, a reconstruction loss from autoencoding an image

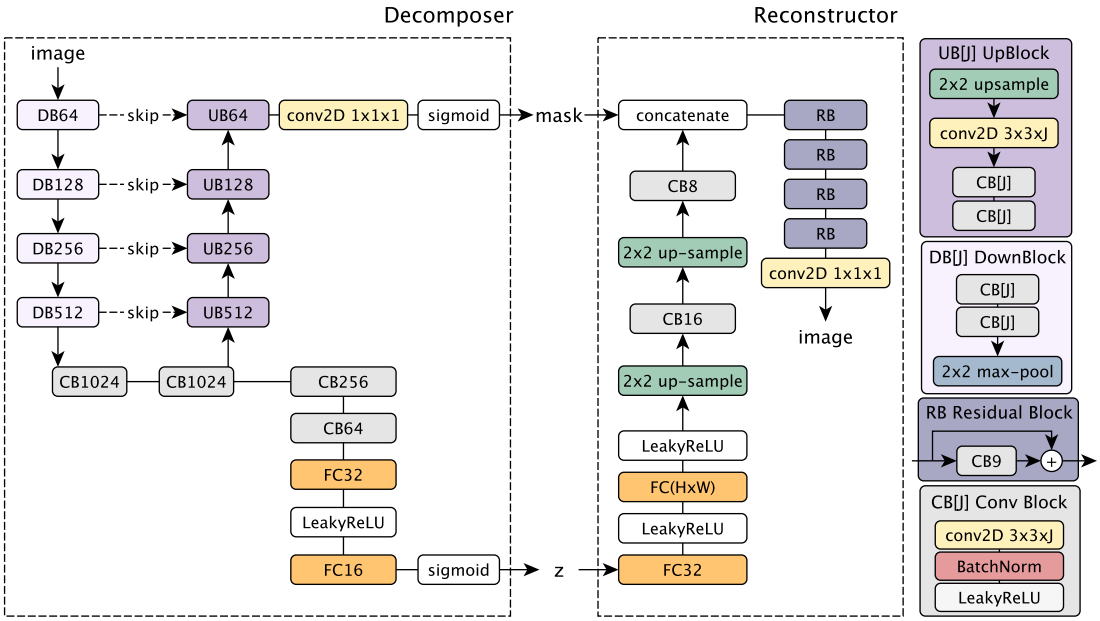


Figure 6.3: Schematic of SMDNet: an image is decomposed as a spatial representation of anatomy (in our case myocardial mask m) and a latent vector z that captures other anatomical and imaging characteristics. Both mask and z are used to reconstruct the input. The model consists of several convolutional (CB) and dense blocks (DB). BatchNormalization and LeakyRelu activations are used throughout.

is defined in Equation 6.1:

$$L_{rec}(f,g) = \mathop{\mathbb{E}}_{x}\left[\|x - g(f(x))\|_1\right]. \tag{6.1}$$

Secondly, two supervised losses when having images with corresponding masks $m \in M$ are defined in Equations 6.2, and 6.3:

$$L_M(f) = \mathop{\mathbb{E}}_{x,m}\left[Dice(m, f_M(x))\right], \tag{6.2}$$

$$L_X(f,g) = \mathop{\mathbb{E}}_{x,m}\left[\|x - g(m, f_Z(x)))\|_1\right]. \tag{6.3}$$

Finally, an adversarial loss using an image discriminator D_X is defined in Equation 6.4, where networks f and g are trained to maximise this objective against an adversarial discriminator trained to minimise it:

$$L^X_{adv}(f,g) = \mathop{\mathbb{E}}_{x}[D_x(g(f(x)))^2 + (D_X(x) - 1)^2]. \tag{6.4}$$

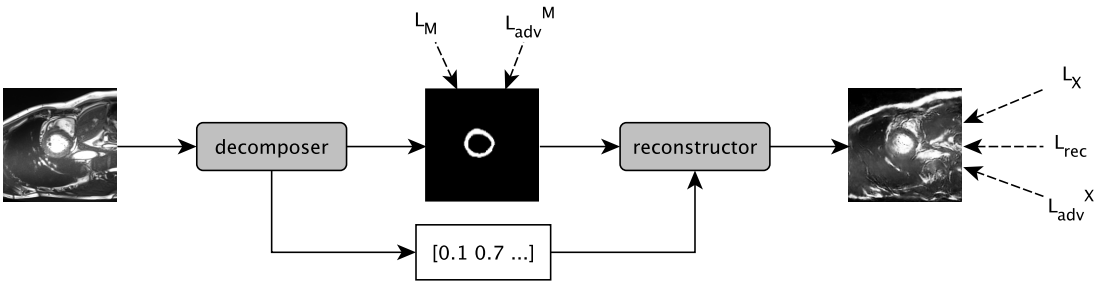


Figure 6.4: An illustration of the training losses of SMDNet.

Similarly, we define an adversarial loss using a mask discriminator D_M in Equation 6.5:

$$L_{adv}^{M}(f) = \mathop{\mathbb{E}}_{x,m}[D_M(f_M(x))^2 + (D_M(m) - 1)^2]. \tag{6.5}$$

Both adversarial losses are based on LeastSquares-GAN [50].

Implementation Details

The decomposer follows a U-Net [13] architecture (see Figure 6.3), and its last layer outputs a segmentation mask of the myocardium via a sigmoid activation function. The model's deep spatial maps contain downsampled image information, which is used to derive the latent vector z through a series of convolutions and fully connected layers, with the final output being passed through a sigmoid so z is bounded. Following this, an architecture with three residual blocks is employed as the reconstructor (see Figure 6.3).

The spatial and continuous representations are not explicitly made independent, so during training the model could still store all information needed for reconstructing the input as low values in the spatial mask, as also observed in [204], since finding a mapping from a spatial representation to an image is easier than combining two sources of information, namely the mask and z. To prevent this, we apply a step function (i.e. a threshold) at the spatial input of the reconstructor to binarise the mask in the forward pass, and encourage the reconstructor to learn a mapping from both binary mask and a vector z to a target image. We store the original values and bypass the step function during back-propagation, and apply the updates to the original non-binary mask. Note that the binarisation of the mask only takes place at the input of the reconstructor network and is not used by the discriminator, in order to encourage the decomposer to produce binary masks, and also train with smoother gradients.

6.4.2 Experimental Setup

6.4.2.1 Data

The experimental evaluation uses 2D cine-MR images that are rescaled to the range [-1,1] from ACDC and QMRI, described in Sections 2.6.2.2 and 2.6.2.3, respectively.

6.4.2.2 Model and Training Details

The overall cost function is a sum of the individual cost functions, which are schematically illustrated in Figure 6.4, and are defined in Equation 6.6:

$$L_{SMDNet}(f,g) = \lambda_1 L_M(f) + \lambda_2 L^M_{adv}(f) + \lambda_3 L_{rec}(f,g) + \lambda_4 L_I(f,g) + \lambda_5 L^X_{adv}(f,g). \quad (6.6)$$

The corresponding loss for images from the unlabelled set does not contain the first and fourth terms. The λ are experimentally set to 10, 10, 1, 10 and 1 respectively. A higher λ value has been selected for cost components that are related to segmentation (L_M and L^M_{adv}), since segmentation is a challenging task. Furthermore, λ_4 has also been set to 10, since L_I uses ground truth segmentations to reconstruct the input, and this is critical for disentanglement of the residual information to the vector component z.

The model is implemented in Keras [202], and trained with Adam [201] with a learning rate of 0.0001. Segmentation results report the test Dice score (Section 3.6) and are obtained through 3-fold cross validation with 70%, 15%, 15% of the images used in training, validation and test splits respectively. SMDNet implementation is available at `https://github.com/agis85/spatial_factorisation`.

6.4.3 Results and Discussion

We demonstrate the proposed decomposition in two ways. Firstly, we show the capability of synthesising new images when combining factors of different slices (see Section 6.4.3.1), and also demonstrate the learned representations by interpolating the latent vector between two images. Secondly in a semi-supervised setting, where we show that we can leverage unlabelled data to increase segmentation accuracy in the few-shot regime (see Section 6.4.3.2).

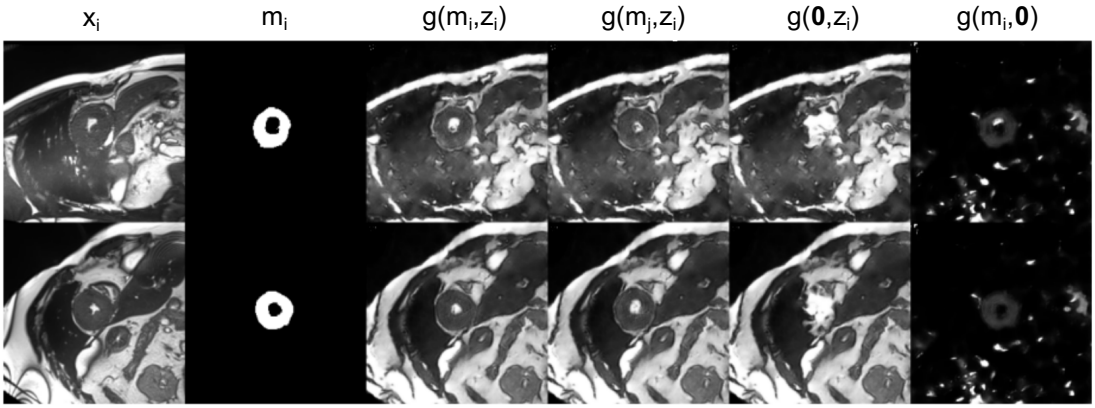


Figure 6.5: Reconstructions using different m_i and z_i combinations for two input images x_1 and x_2 (one per row), respectively. From left to right the columns contain the following: predicted segmentation masks m_1 and m_2; reconstructions $g(m_1, z_1)$ and $g(m_2, z_2)$; synthetic images $g(m_2, z_1)$ and $g(m_1, z_2)$ by mixing masks and vectors; synthetic images $g(\mathbf{0}, z_1)$ and $g(\mathbf{0}, z_2)$ by using a mask of zeros, which has the effect of producing cardiac images without myocardium; finally, synthetic images $g(m_1, \mathbf{0})$ and $g(m_2, \mathbf{0})$ of only the myocardium.

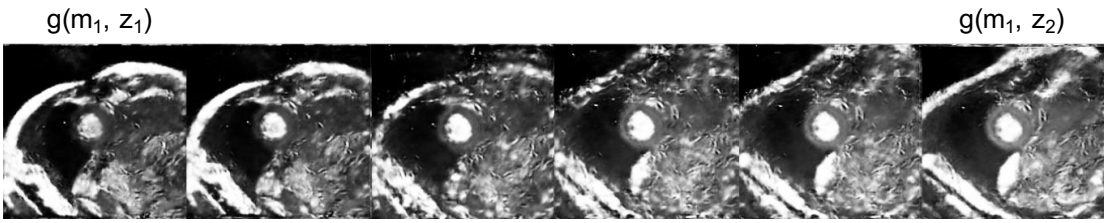


Figure 6.6: Reconstructions when using a fixed mask m_1 and interpolating between two vectors z_1 and z_2.

6.4.3.1 Latent Space Arithmetic

As a demonstration of the learned representation, Figure 6.5 shows reconstructions of input images from the training set using different combinations of masks and z components. The first three columns show the original input with the predicted mask and the input's reconstruction. Next, we take the spatial representation m_j from one image and combine it with the z_i component of the other image, and vice versa. As shown in the figure (4th column) the intensities and the anatomy around the myocardium remain unchanged, but the myocardial shape and position, which are encoded in the mask, change to that of the second image. The final two columns show reconstructions using a null mask (i.e. $m_i = \mathbf{0}$) and the correct z_i in 5th column,

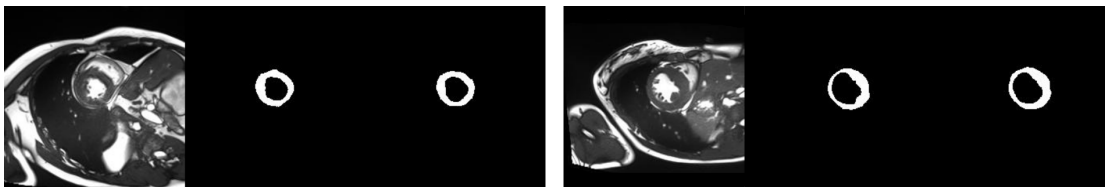


Figure 6.7: Two examples of segmentation performance: input, prediction and ground truth.

or using the original mask with a $z_i = \mathbf{0}$ in the 6th column. In the first case, the produced image does not contain myocardium, whereas in the second case the image contains only myocardium and no other anatomical or MR characteristics.

Moreover, we qualitatively evaluate the smoothness of the residual representation with synthetic images presented in Figure 6.6, which demonstrates how the underlying anatomy changes slowly from the left-most image to the right-most image. The synthetic images are produced by using a fixed myocardium mask, and by interpolating between two vectors z_1 and z_2, that are obtained from the real input images.

6.4.3.2 Semi-supervised Results

The utility of the disentangled representation becomes evident in semi-supervised learning. Qualitatively in Figure 6.7 we can see that our method closely follows ground truth segmentation masks (example from ACDC held-out test set) when trained with full supervision. Further-

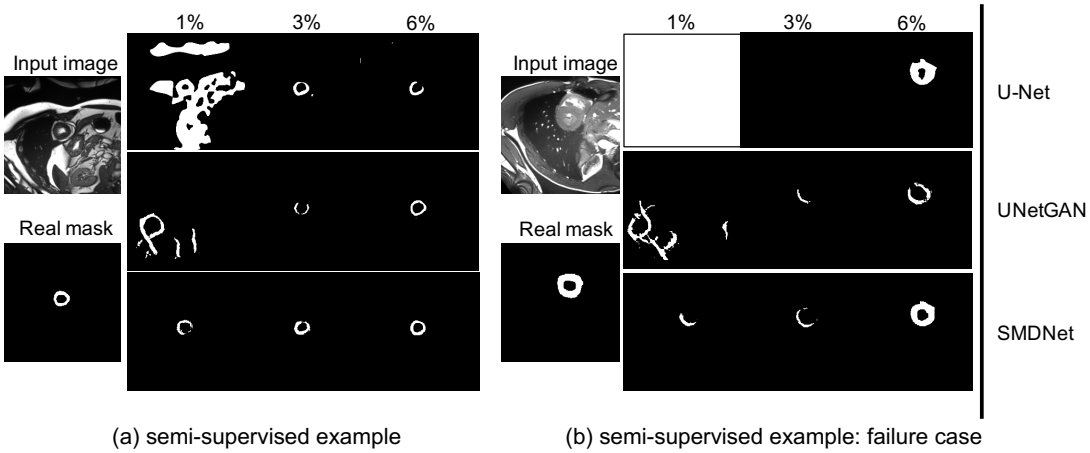


Figure 6.8: Example segmentation masks produced by U-Net, GAN, and SMDNet trained in ACDC on low fractions of labelled data.

Labels	100%	50%	25%	12.5%	6%	3%	1.5%
U-Net	81.7_{08}	80.0_{08}	78.2_{09}	65.7_{19}	58.1_{17}	35.6_{23}	02.6_{01}
self-train	79.2_{10}	71.1_{14}	52.0_{27}	54.2_{21}	48.2_{19}	31.1_{14}	03.9_{02}
UNetGAN	$\mathbf{82.6}_{07}$	77.2_{11}	$\mathbf{78.7}_{09}$	72.7_{10}	64.8_{13}	36.5_{32}	08.0_{06}
SMDNet	82.2_{08}	$\mathbf{81.4}_{08}$	77.1_{09}	$\mathbf{76.7}_{09}$	$\mathbf{73.1}_{12}$	$\mathbf{67.8}_{14}$	$\mathbf{41.5}_{13}$

Table 6.1: Dice scores (%) and standard deviations of myocardium on ACDC data. The models are trained with 1200 unlabelled images, and different proportions of labelled data shown in the top row. The masks used for adversarial training do not correspond to any training images. Best results are shown in bold font.

more, Figure 6.8 shows two segmentation examples produced by U-Net, UNetGAN, SMDNet when trained on ACDC for 1%, 3% and 6% labelled images. At 1%, which corresponds to 11 images, the U-Net collapses and cannot produce a good segmentation, whereas the UNetGAN produces a mask that looks circular to partially satisfy the shape constraint of the myocardium. Even in the failure case of Figure 6.8b, SMDNet results are more consistent, although the myocardium is under-segmented.

To assess our performance quantitatively we train a variety of setups varying the number of labelled training images whilst keeping the unlabelled fixed (in both ACDC and QMRI cases). We train SMDNet and the benchmarks (U-Net, self-train, and UNetGAN), and report results on held-out test sets in Tables 6.1 and 6.2 for the two datasets respectively. We can see that even when the number of labelled images is very low, our method is able to achieve segmentation accuracy considerably higher than the other two methods. As the number of labelled images increases, all models perform comparably good.

An extension can be considered, in which the spatial factor is a generic factor of the anatomy and does not restrict to the myocardium. This offers many benefits, since it enables multi-class segmentation, and also multi-task learning of further anatomical tasks. This extension is presented in Section 6.5 below.

Labels	100%	50%	25%	12.5%
U-Net	68.6_{10}	68.1_{09}	44.1_{15}	36.8_{17}
self-train	70.2_{09}	50.1_{22}	23.0_{29}	06.1_{07}
UNetGAN	$\mathbf{79.5}_{06}$	75.6_{07}	58.0_{12}	06.1_{06}
SMDNet	79.4_{04}	$\mathbf{77.2}_{07}$	$\mathbf{68.6}_{14}$	$\mathbf{42.4}_{14}$

Table 6.2: Dice scores (%) and standard deviations of myocardium on QMRI data. The models are trained with 1200 unlabelled images, and different proportions of labelled data shown in the top row. The masks used for adversarial training do not correspond to any training images. Best results are shown in bold font.

6.5 Spatial Disentanglement Network

Section 6.4 presented a myocardial decomposition method for cardiac MR images. Learning a decomposition of data into a spatial content factor and a non-spatial style factor has been a focus of recent research in computer vision [25, 131] with the aim being to achieve diversity in style transfer between domains. However, no consideration has been taken regarding the semantics and the precision of the spatial factor. This is crucial in medical analysis tasks in order to be able to extract quantifiable information directly from the spatial factor. In our previous work of Section 6.4, we aimed to precisely address the need for interpretable semantics by explicitly enforcing the spatial factor to be a binary myocardial segmentation. However, since the spatial factor is a segmentation mask of only the myocardium, remaining anatomies must be encoded in the non-spatial factor, which violates the concept of explicit factorisation into anatomical and modality factors.

In this section instead, we propose *Spatial Disentanglement Network* (SDNet), schematic shown in Figure 6.9, that learns a disentangled representation of medical images consisting of a spatial map that semantically represents the anatomy, and a non-spatial latent vector containing image modality information.

The anatomy is modelled as a multi-channel feature map, where each channel represents different anatomical substructures (e.g. myocardium, left and right ventricles). This spatial representation is categorical with each pixel necessarily belonging to exactly one channel. This strong restriction prevents the binary maps from encoding modality information, encouraging the anatomy factors to be modality-agnostic (invariant), and further promotes factorisation of

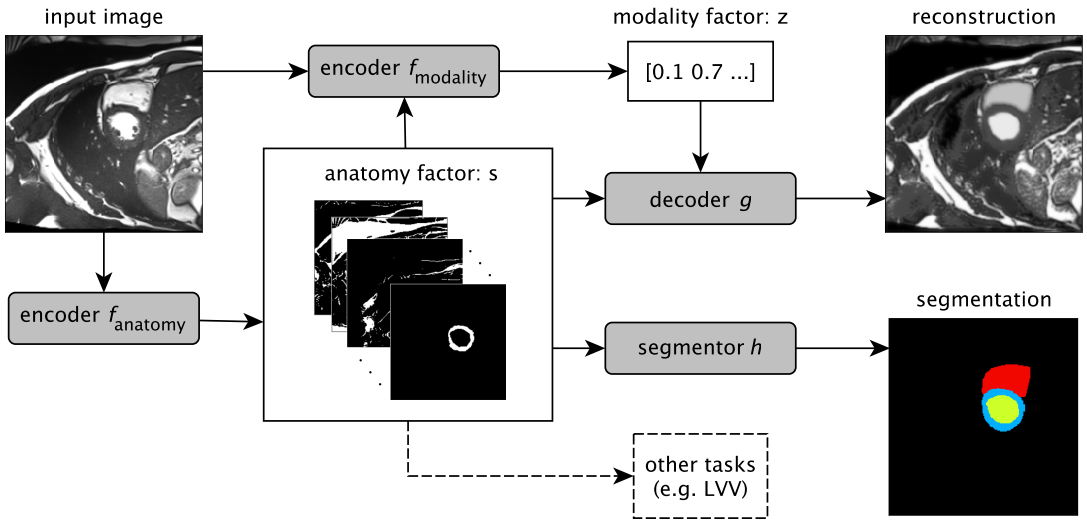


Figure 6.9: A schematic overview of the proposed model. An input image is first encoded to a multi-channel spatial representation, the anatomy factor s, using an anatomy encoder $f_{anatomy}$. Then s can be used as an input to a segmentation network h to produce a multi-class segmentation mask, (or some other task specific network). The factor s along with the input image are used by a modality encoder $f_{modality}$ to produce a latent vector z representing the imaging modality. The two representations s and z are combined to reconstruct the input image through the decoder network g.

the subject's anatomy into meaningful topological regions.

On the other hand, the non-spatial factor contains modality-specific information, in particular the distribution of intensities of the spatial regions. We encode the image intensities into a smooth latent space, using a Variational Autoencoder (VAE) loss, such that nearby values in this space correspond to neighbouring values in the intensity space.

Finally, since the representation should retain most of the required information about the input (albeit in two factors), image reconstructions are possible by combining both factors.

6.5.1 Materials and Methods

Overall, our proposed model can be considered as an autoencoder, which takes as input a 2D volume slice $x \in X$, where $X \subset \mathbb{R}^{H \times W}$ is the set of all images in the data, with H and W being the image's height and width respectively. The model generates a reconstruction through

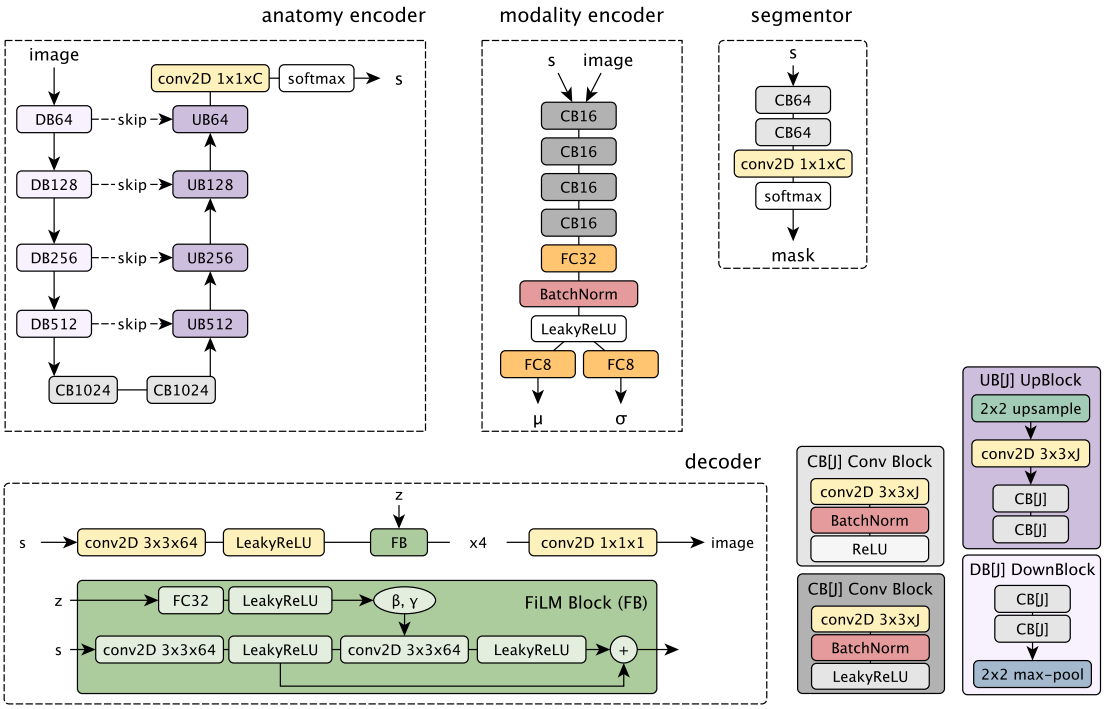
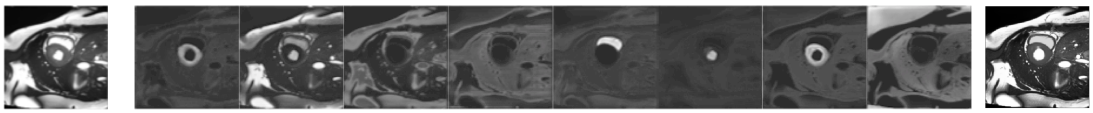


Figure 6.10: The architectures of the four networks that make up SDNet. The anatomy encoder is a standard U-Net [13] that produces a spatial anatomical representation s. The modality encoder is a convolutional network (except for a fully connected final layer) that produces the mean μ and standard deviation σ of a Gaussian distribution, used to sample the modality representation z. The segmentor is a small fully convolutional network that produces the final segmentation prediction of a multi-class mask (with V classes) given s. Finally the decoder produces a reconstruction of the input image from s with its output modulated by z through FiLM [14]. The anatomy factor's channels parameter C, the modality factor's size n_z, and the number of segmentation classes V depend on the specific task and are detailed in the main text.

an intermediate disentangled representation. The disentangled representation is comprised of a multi-channel spatial map (a tensor) $s \in S \coloneqq \{0,1\}^{H \times W \times C}$, where C is the number of channels, and a multi-dimensional continuous vector factor $z \in Z \coloneqq \mathbb{R}^{n_z}$, where n_z is the number of dimensions. These are generated respectively by two encoders, modelled as convolutional neural networks, $f_{anatomy}$ and $f_{modality}$. The two representations are combined by a decoder g to reconstruct the input. In addition to the reconstruction cost, explicit supervision can be given in the form of auxiliary tasks, for example with a segmentation task using a network h, or with a regression task as we will demonstrate in Section 6.5.3.2. A schematic of our model can be seen in Figure 6.9 and the detailed architectures of each network are shown in Figure 6.10.

(a) Input, anatomical representation with binary thresholding, and reconstruction image.



(b) Input, anatomical representation with no binary thresholding, and reconstruction image.

Figure 6.11: (a) Example of a spatial representation, expressed as a multi-channel binary map. Some channels represent defined anatomical parts such as the myocardium or the left ventricle, and others the remaining anatomy required to describe the input image on the left. Observe how sparse most of the informative channels are. (b) Spatial representation with no thresholding applied. Each channel of the spatial map, also captures the intensity signal in different gray level variations and is not sparse, in contrast to Figure 6.11a. This may hinder an anatomical separation. Note that no specific channel ordering is imposed and thus the anatomical parts can appear in different order in the anatomical representations across experiments.

6.5.1.1 Input Decomposition

The decomposition process yields representations for the anatomy and the modality characteristics of medical images and is achieved by two dedicated neural networks. Whilst a decomposition could also be performed with a single neural network with two separate outputs and shared layer components, as done in our previous work (Section 6.4), we found that by using two separate networks, as also done in [25, 131], we can control more easily the information captured by each factor, and stabilise the behaviour of each encoder during training.

Anatomical Representation

The anatomy encoder is a fully convolutional neural network that maps 2D images to spatial representations, $f_{anatomy} : X \rightarrow S$. We use a U-Net [13] architecture, containing downsampling and upsampling paths with skip connections between feature maps of the same size, allowing effective fusion of important local and non-local information.

The spatial representation is a feature map consisting of a number of binary channels of the same spatial dimensions as the input image, that is $s \in \{0,1\}^{H \times W \times C} s.t. \sum_{c=1}^{C} s_{h,w,c} = 1$

$\forall h \in \{1, \ldots, H\}, w \in \{1, \ldots, W\}$, where C is the number of channels. Some channels contain individual anatomical (cardiac) sub-structures, while the other structures, necessary for reconstruction, are freely dispersed in the remaining channels. Figure 6.11a shows an example of a spatial representation, where the myocardium, the left and the right ventricle are clearly visible, and the remaining channels contain the surrounding image structures (albeit more mixed and not anatomically distinct).

The spatial representation is derived using a softmax activation function to force each pixel to have activations that sum to one across the channels. Since softmax functions encode continuous distributions, we binarise the anatomical representation via the operator $s \mapsto \lfloor s + 0.5 \rfloor$, which acts as a threshold for the pixel values of the spatial variables in the forward pass. During back-propagation the step function is bypassed and updates are applied to the original non-binary representation, as in the straight-through operator [205].

Thresholding s is integral to the model design and offers two advantages. Firstly, it reduces the capacity of the spatial factor, encouraging it to be a representation of only the anatomy, while preventing encoding modality information. Secondly, it enforces a factorisation of the spatial factor in distinct channels, as each pixel can only be active on one channel. To illustrate the importance of this binarisation, an example of a non-thresholded spatial factor is shown in Figure 6.11b. Observe, that the channels of s are not sparse with gray level variations now evident. Image intensities are encoded spatially, using different grayscale values, allowing good reconstructions to be achieved without a modality factor, which we explicitly want to avoid.

Modality Representation

Given samples of the data $x \in X$ with their corresponding $s \in S$ (deterministically obtained by $f_{anatomy}$), we learn the posterior distribution of latent factors $z \in Z \coloneqq \mathbb{R}^{n_z}$, $q(z|x, s)$.

Learning this posterior distribution follows the VAE principle [44]. In brief a VAE learns a low dimensional latent space, such that the learnt latent representations match a prior distribution that is set to be an isotropic multivariate Gaussian $p(z) = \mathcal{N}(0, 1)$. A VAE consists of an encoder and a decoder. The encoder, given an input, predicts the parameters of a Gaussian distribution (with diagonal co-variance matrix). This distribution is then sampled, using the reparameterisation trick to allow learning through back propagation, and the resulting sample is fed through the decoder to reconstruct the input. VAEs are trained to minimise a reconstruction

error and the KL divergence of the estimated Gaussian distribution $q(z|x,s)$ from the unit Gaussian $p(z)$ of Equation 6.7:

$$L_{KL}(f_{anatomy}, f_{modality}) = \mathop{\mathbb{E}}_{x}\left[KL(q(z|x,s)\|p(z))\right], \tag{6.7}$$

where $KL(p\|q) = \int p(z) log \frac{p(z)}{q(z|x,f_{anatomy}(x))} dz$. Once trained, sampling a vector from the unit Gaussian over the latent space and passing it through the decoder approximates sampling from the data, i.e. the decoder can be used as a generative model.

The posterior distribution is modelled with a stochastic encoder (this is analogous to the VAE encoder) as a convolutional network, which encodes the image modality, $f_{modality} : X \times S \rightarrow Z$. Specifically, the stochasticity of the encoder (for a sample x and its anatomy factor s) is achieved as in the VAE formulation as follows: $f_{modality}(x,s)$ produces first the mean and standard deviation for an n_z dimensional Gaussian, which is then sampled to yield the final z.

6.5.1.2 Segmentation

One important task for the model is to infer segmentation masks $m \in M \coloneqq \{0,1\}^{H \times W \times V}$, where V is the number of anatomical segmentation categories in the training dataset, out of the spatial representation. This is an integral part of the training process because it also defines the anatomical structures that will be extracted from the image. The segmentation network[1] is a fully convolutional network consisting of two convolutional blocks followed by a final 1×1 convolution layer (see Figure 6.10), with the goal of refining the anatomy present in the spatial maps and produce the final segmentation masks, $h : S \rightarrow M$.

When labelled data are available, a supervised cost is employed that is based on a differentiable Dice loss [206] between a real segmentation mask m of an image sample x and its predicted segmentation $h(f_{anatomy}(x))$, described in Equation 6.8:

$$L_{segm}(f_{anatomy}, h) = 1 - 2 \times \mathop{\mathbb{E}}_{x,m}\left[\frac{\sum_{h,w,l}(m_{h,w,l} \times h(f_{anatomy}(x))_{h,w,l}) + \epsilon}{\sum_{h,w,l}(m_{h,w,l} + h(f_{anatomy}(x))_{h,w,l}) + \epsilon}\right], \tag{6.8}$$

where the added small constant ϵ prevents division by 0. In a semi-supervised scenario, where

[1]Experimental results showed that having an additional segmentor network, instead of enforcing our spatial representation to contain the exact segmentation masks, improves the training stability of our method. Furthermore, it offers flexibility in that the same anatomical representation can be used for multiple tasks, such as in segmentation and the calculation of the left ventricular volume.

there are images with no corresponding segmentations, an adversarial loss is defined in Equation 6.9, using a discriminator over masks D_M, based on LeastSquares-GAN [50]. Networks $f_{anatomy}$ and h are trained to maximise the adversarial objective, against D_M which is trained to minimise it:

$$L_{adv}(f_{anatomy}, h) = \mathop{\mathbb{E}}_{x,m} \left[D_M(h(f_{anatomy}(x)))^2 + (D_M(m) - 1)^2 \right]. \tag{6.9}$$

The architecture of the discriminator is based on DCGAN discriminator [48], without Batch Normalization.

6.5.1.3 Image Reconstruction

The two factors are combined by a decoder network g to generate an image $y \in Y \coloneqq \mathbb{R}^{H \times W \times 1}$ with the anatomical characteristics specified by s and the imaging characteristics specified by z, $g : S \times Z \rightarrow Y$. The fusion of the two factors acts as an inpainting mechanism where the information stored in z, is used to derive the image signal intensities that will be used on the anatomical structures, stored in s.

The reconstruction is achieved by a decoder convolutional network conditioned with four FiLM [14] layers. This general purpose conditioning method learns scale and offset parameters for each feature-map channel within a convolutional architecture. Thus, an affine transformation (one per channel) learnt from the conditioning input is applied.

Here, a network of two fully connected layers (see Figure 6.10) maps z to the scale and offset values γ and β for each intermediate feature map F_c of the decoder. Each channel of F_c is modulated based on c pairs γ_c and β_c as follows: $FiLM(F_c|\gamma_c, \beta_c) = \gamma_c \odot F_c + \beta_c$, where element-wise multiplication ($\odot$) and addition are both broadcast over the spatial dimensions. The decoder and FiLM parameters are learnt through the reconstruction of the input images using the MAE, defined in Equation 6.10:

$$L_{rec}(f_{anatomy}, f_{modaliy}, g) = \mathop{\mathbb{E}}_{x} \left[\| x - g(f_{anatomy}(x), f_{modality}(x, f_{anatomy}(x))) \|_1 \right]. \tag{6.10}$$

The design of the decoding process restricts the type of information stored in z to only affect the intensities of the produced image. This is important in the disentangling process as it pushes z to not contain spatial anatomical information.

The decoder can also be interpreted as a conditional generative model, where different samples of z conditioned on a given s generate images of the same anatomical properties, but with different appearances. The reconstruction process is the opposite of the decomposition process, i.e. it learns the dependencies between the two factors in order to produce a realistic output.

Modality Factor Reconstruction

A common problem when training VAE is posterior collapse: a degenerate condition where the decoder is ignoring some factors. In this case, even though the reconstruction is accurate, not all data variation is captured in the underlying factors.

In our model posterior collapse manifests when some modality information is spatially encoded within the anatomy factor.[2] To overcome this we use a z reconstruction cost (Equation 6.11), according to which an image y produced by a random z sample should produce the same modality factor when (re-)encoded,

$$L^z_{rec}(f_{anatomy}, f_{modality}, g) = \mathop{\mathbb{E}}_{z,x} \left[\| z - f_{modality}(y, f_{anatomy}(y)) \|_1 \right]. \tag{6.11}$$

The faithful reconstruction of the modality factor z penalises the VAE for ignoring dimensions of the latent distribution and encourages each encoded image to produce a low variance Gaussian. This is in tension with the KL divergence cost which is optimal when the produced distribution is a spherical Gaussian of zero mean and unit variance. A perfect score of the KL divergence results in all samples producing the same distribution over z, and thus the samples are indistinguishable from each other based on z. Without L^z_{rec}, the overall cost function can be minimised if imaging information is encoded in s, thus resulting in posterior collapse. Reconstructing the modality factor prevents this, and results in an equilibrium where a good reconstruction is possible only with the use of both factors.

6.5.2 Experimental Setup

6.5.2.1 Data

Experiments use 2D images from four datasets, that are normalised to the range [-1, 1].

[2] Note that while using FiLM prevents z from encoding spatial information, it does not prevent the case of posterior collapse i.e. that s encodes (all or part of) the modality information.

(a) For the semi-supervised segmentation experiment (Section 6.5.3.1) and the latent space arithmetic (Section 6.5.3.5) we use the ACDC dataset, described in Section 2.6.2.2.

(b) We also use data acquired at Edinburgh Imaging Facility QMRI (Section 2.6.2.3) for the semi-supervised segmentation and multi-task experiments of Sections 6.5.3.1 and 6.5.3.2 respectively.

(c) Finally, we use cine-MR and CP-BOLD images from the BOLD dataset (Section 2.6.2.4 to further evaluate modality estimation (Section 6.5.3.4).

6.5.2.2 Model and Training Details

The overall cost function is a composition of the individual costs of each of the model's components and is defined in Equation 6.12:

$$L_{SDNet} = \lambda_1 L_{KL} + \lambda_2 L_{segm} + \lambda_3 L_{adv} + \lambda_4 L_{rec} + \lambda_5 L^z_{rec}. \tag{6.12}$$

The λ parameters are experimentally set to values: λ_1=0.01, λ_2=10, λ_3=10, λ_4=1, λ_5=1. We opt for a lower λ_1 to prevent posterior collapse in the decoder (which would ignore z) and adopt the value from [129] that also trains a VAE for modelling intensity variability. Separating the anatomy into segmentation masks is a difficult task, and is also in tension with the reconstruction process which pushes parts with similar intensities to be in the same channels. This motivates our decision in increasing the values of the segmentation hyperparameters λ_2 and λ_3. The remaining λ_4 and λ_5 are set to the default value of 1, such that the errors are in the same value range as the errors of the previous loss components.

We set the dimension of the modality factor n_z=8 as in [129] across all datasets. We also set the number of channels of the spatial factor to C=8 for ACDC and QMRI and increase to C=16 for MM-WHS, to support the increased number of segmented regions (7 in MM-WHS) and the fact that CT and MR data have different contrasts and viewpoints. This additional flexibility allows the network to use some channels of s for common information across the two modalities (MR and CT) and some for unique (not common) information.

We train using Adam [201] with a learning rate of 0.0001. We used a batch size of 4 and an early stopping criterion based on the segmentation cost of a validation set. All code was developed in Keras [202]. The quantitative results of Section 6.5.3 are obtained through 3-fold

cross validation, where each split contains a proportion of the total volumes of 70%, 15% and 15% corresponding to training, validation and test sets. SDNet implementation is available at `https://github.com/agis85/anatomy_modality_decomposition`.

6.5.3 Results

We here present and discuss quantitative and qualitative results of our method in various experimental scenarios. Initially, multi-class semi-supervised segmentation is evaluated in Section 6.5.3.1. Subsequently, Section 6.5.3.2 demonstrates multi-task learning with the addition of a regression task in the training objectives. In Section 6.5.3.3, SDNet is evaluated in a multi-modal scenario by concurrently segmenting MR and CT data. In Section 6.5.3.4 we investigate whether the modality factor z captures multimodal information. Finally, Section 6.5.3.5 demonstrates properties of the factorisation using latent space arithmetic, in order to show how z and s interact to reconstruct images.

6.5.3.1 Semi-supervised Segmentation

We evaluate the utility of our method in a semi-supervised experiment, in which we combine labelled images with a pool of unlabelled images to achieve multi-class semi-supervised segmentation. Specifically, we explore the sensitivity of SDNet and the baselines of Section 6.3 to the number of labelled examples, by training with various numbers of labelled images. Our objective is to show that we can achieve comparable results to a fully supervised network using fewer annotations.

To simulate a more realistic clinical scenario, sampling of the labelled images does not happen over the full image pool, but at a subject level: initially, a number of subjects are sampled, and then all images of these subjects constitute the labelled dataset. The number of unlabelled images is fixed and set equal to 1200 images: these are sampled at random from all subjects of the training set and from cardiac phases other than End Systole (ES) and End Diastole (ED) (for which no ground truth masks exist). The real segmentation masks used to train the mask discriminator are taken from the set of image-mask pairs from the same dataset.

In order to test the generalisability of all methods to different types of images, we use two cine-MR datasets: ACDC which contains masks of the LV, MYO and RV; and QMRI which contains masks of the LV and MYO. Spatial augmentations by rotating inputs up to 90° are applied

Labels	U-Net				UNetGAN				self-train				SDNet			
	MYO	LV	RV	avg	MYO	LV	RV	avg	MYO	LV	RV	avg	MYO	LV	RV	avg
100%	83_{7}	88_{6}	79_{10}	$\mathbf{85_{7}}$	82_{6}	87_{6}	75_{8}	83_{5}	$\mathbf{84_{7}}$	$\mathbf{89_{5}}$	$\mathbf{82_{8}}$	$\mathbf{85_{6}}$	$\mathbf{84_{5}}$	88_{4}	78_{8}	84_{5}
50%	$\mathbf{83_{7}}$	$\mathbf{87_{7}}$	$\mathbf{79_{10}}$	$\mathbf{85_{7}}$	81_{7}	86_{6}	75_{10}	82_{7}	80_{10}	85_{10}	78_{11}	82_{8}	$\mathbf{83_{6}}$	$\mathbf{87_{7}}$	77_{9}	83_{6}
25%	77_{9}	82_{9}	67_{14}	75_{11}	78_{9}	$\mathbf{85_{8}}$	72_{11}	79_{8}	76_{13}	$\mathbf{85_{10}}$	70_{15}	78_{11}	$\mathbf{80^{*}_{7}}$	$\mathbf{85_{6}}$	$\mathbf{73_{11}}$	$\mathbf{81^{*}_{6}}$
12.5%	71_{13}	80_{13}	61_{17}	70_{13}	78_{8}	$\mathbf{85_{6}}$	$\mathbf{69_{13}}$	79_{8}	63_{17}	77_{13}	57_{21}	67_{15}	$\mathbf{79_{8}}$	$\mathbf{85_{7}}$	$\mathbf{69_{13}}$	$\mathbf{80_{8}}$
6%	63_{12}	76_{13}	56_{22}	65_{13}	75_{11}	81_{11}	69_{13}	75_{12}	46_{27}	59_{23}	34_{18}	47_{23}	$\mathbf{77_{9}}$	$\mathbf{83_{10}}$	$\mathbf{71_{12}}$	$\mathbf{78^{*}_{9}}$
3%	55_{19}	66_{20}	46_{20}	52_{18}	73_{32}	79_{10}	67_{14}	75_{10}	20_{15}	35_{20}	22_{14}	24_{15}	$\mathbf{76^{*}_{7}}$	$\mathbf{82^{*}_{8}}$	$\mathbf{68_{14}}$	$\mathbf{77^{*}_{8}}$
1.5%	26_{19}	33_{21}	35_{17}	21_{19}	67_{21}	78_{11}	63_{12}	67_{12}	11_{10}	19_{14}	25_{12}	16_{11}	$\mathbf{70_{12}}$	$\mathbf{77_{13}}$	$\mathbf{64_{15}}$	$\mathbf{73^{*}_{12}}$

Table 6.3: Dice score (%) on ACDC for MYO, LV, RV, and average. Standard deviations are shown as subscripts. The models are trained with 1200 unlabelled and different fractions of labelled images (each one corresponding to a proportion of selected subjects). For each of the three components and the average separately, the best result is shown in bold font and an asterisk indicates statistical significance at the 5% level compared to the second best method in the same row/component.

to experiments using ACDC data to better simulate the orientation variability of the dataset. No augmentations are applied in experiments using QMRI data since all images maintain a canonical orientation.[3] No further augmentations have been performed to fairly compare the effect of the different methods.

We present the average cross-validation Dice score (Section 3.6) on held out test sets across all labels, as well as the Dice score for each label separately, and the corresponding standard deviations. Note that images from a given subject can only be present in exactly one of the training, validation or test sets. Table 6.3 contains the ACDC results for all labels, MYO, LV and RV respectively, and Table 6.4 contains the QMRI results for all labels, MYO, and LV respectively. The test set for each fold contains 280 images of ED and ES phases, belonging to 15 subjects for ACDC, and 35 images of the ED phase belonging to 4 subjects for QMRI. The best results are shown in bold font, and an asterisk indicates statistical significance at the 5% level, compared to the second best result, computed using a paired t-test. In both tables the

[3]Using data that present different (non-canonical) orientations is possible to affect segmentation performance, since features extracted by neural networks are not rotation invariant and thus might be biased to the training data. This can be solved with appropriate data augmentation.

Labels	U-Net			UNetGAN			self-train			SDNet		
	MYO	LV	avg	MYO	LV	avg	MYO	LV	avg	MYO	LV	avg
100%	72_{9}	90_{6}	83_{7}	$\mathbf{75_{7}}$	$\mathbf{93_{3}}$	$\mathbf{86_{4}}$	$\mathbf{75_{9}}$	92_{5}	$\mathbf{86_{7}}$	$\mathbf{75_{6}}$	$\mathbf{93_{4}}$	$\mathbf{86_{4}}$
50%	72_{15}	82_{18}	74_{15}	71_{9}	86_{7}	83_{5}	62_{11}	88_{9}	79_{9}	$\mathbf{73_{6}}$	$\mathbf{90_{5}}$	$\mathbf{84_{5}}$
25%	54_{14}	80_{9}	69_{10}	$\mathbf{68_{7}}$	86_{7}	$\mathbf{81_{5}}$	36_{22}	56_{29}	49_{26}	66_{7}	$\mathbf{88_{7}}$	80_{8}
12.5%	52_{11}	81_{6}	65_{7}	68_{8}	88_{6}	79_{7}	42_{16}	64_{14}	58_{14}	$\mathbf{67_{9}}$	$\mathbf{88_{6}}$	$\mathbf{80_{7}}$
6%	21_{14}	43_{28}	43_{20}	64_{9}	$84_{1}0$	75_{10}	8_{6}	21_{11}	13_{7}	$\mathbf{65_{7}}$	$\mathbf{87_{10}}$	$\mathbf{79_{5}}$

Table 6.4: Dice score (%) on QMRI for MYO, LV, and average. Standard deviations are shown as subscripts. The models are trained with 1200 unlabelled and different fractions of labelled images (each one corresponding to a proportion of selected subjects). For each of the two components and the average separately, the best result is shown in bold font and an asterisk indicates statistical significance at the 5% level compared to the second best method in the same row/component.

lowest amount of labelled data (1.5% for Table 6.3 and 6% for Table 6.4) correspond to images selected from one subject. Segmentation examples for ACDC data using different number of labelled images are shown in Figure 6.12, where different colours are used for the different segmentation classes.

For both datasets, when the number of annotated images is high, then all methods perform equally well, although our method achieves the lowest variance. In Table 6.3 the performance of the supervised (U-Net) and self-trained methods decreases when the number of annotated images reduces below 12.5%, since the limited annotations are not sufficiently representative of the data. When using data from one or two subjects, these two methods which mostly rely on supervision fail with a Dice score below 55%. On the other hand, even when the number of labelled images is small, adversarial training used by SDNet and UNetGAN helps maintaining a good performance. The reconstruction cost used by our method further regularises training and consistently produces more accurate results, with Dice scores equal to 73%, 77% and 78% for 1.5%, 3% and 6% labels respectively, that are also significantly better, with p-values 0.0006, 0.02, and 0.002, in a paired t-test.

It is interesting to compare the performance of SDNet with our previous work (Section 6.4). We therefore modify our previous model for multi-class segmentation and repeat the experiment for

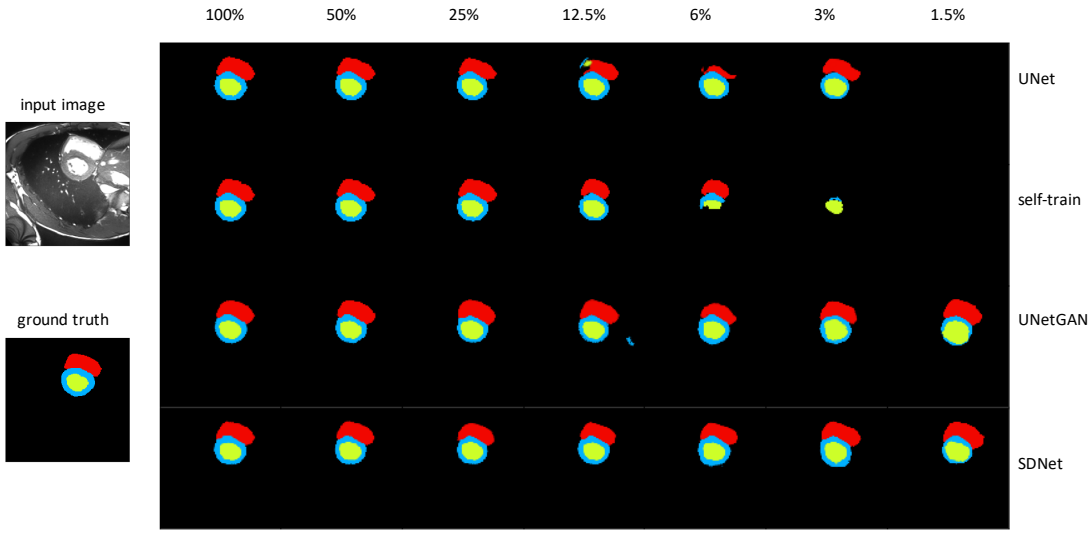


Figure 6.12: Segmentation example for different numbers of labelled images from the ACDC dataset. Blue, green and red show the models prediction for MYO, LV and RV respectively.

the ACDC dataset. We compute the Dice scores and standard deviations for 100%, 50%, 25%, 12.5%, 6%, 3%, and 1.5% of labelled data to be respectively $79 \pm 7\%$, $75 \pm 8\%$, $79 \pm 7\%$, $77 \pm 10\%$, $75 \pm 9\%$, $66 \pm 15\%$, and $59 \pm 13\%$. Comparing with the results of Table 6.3, SDNet significantly outperforms our previous model (at the 5% level, paired t-test).

On the smaller QMRI dataset, the segmentation results are seen in Table 6.4, and correspond to two masks instead of three. When using annotated images from just a single subject (corresponding to 6% of the data the lowest possible), the performance of the supervised method reduces by almost 50% compared to when using the full dataset. SDNet and UNetGAN both maintain a good performance of 75% and 79%, with no significant differences between them.

6.5.3.2 Left Ventricular Volume

It is common for clinicians to not manually annotate all endocardium and epicardium contours for all patients if it is not necessary. Rather, a mixture of annotations and other metrics of interest will be saved at the end of the study in the electronic health record. For example, we can have a scenario with images of some patients that contain myocardium segmentations and some images with the value of their left ventricular volume. Here we test our model in such a multi-task scenario and show that we can benefit from such auxiliary and mixed annotations. We will evaluate, firstly whether our model is capable of predicting a secondary output related

to the anatomy (the volume of the left ventricle), and secondly whether this secondary task improves the performance of the main segmentation task.

Using the QMRI dataset, we first calculate the ground truth left ventricular volume (LVV) for each patient as follows: for each 2D slice, we first sum the pixels of the left ventricular cavity, then multiply this sum with the pixel resolution to get the corresponding area and then multiply the result with the slice thickness to get the volume occupied by each slice. The final volume is the sum of all individual slice volumes.

Predicting the LVV as another output of SDNet follows a similar process to the one used to calculate the ground truth values. We design a small neural network consisting of two convolutional layers (each having a $3 \times 3 \times 16$ kernel followed by a ReLU activation), and two fully connected layers of 16 and 1 neurons respectively, both followed by a ReLU activation. This network regresses the sum of the pixels of the left ventricle, taking as input the spatial representation. The predicted sum can then be used to calculate the LVV offline.

Using a pre-trained model of labelled images corresponding to one subject (last row in Table 6.4 with 6% labels), we fine-tune the whole model whilst training the area regressor using ground truth values from 17 subjects. We find the average LVV over the test volumes equal to $138.57mL$ (standard deviation of 8.8), and the ground truth LVV equal to $139.23mL$ (standard deviation of 2.26), with no statistical difference between them in a paired t-test. Both measurements agree with the normal LVV values for ED cardiac phases, which was reported as $143mL$ in a large population study [207]. The multi-task objective used to fine-tune the whole model also benefits test segmentation accuracy, which is raised from 75.6% to 83.2% (statistically significant at the 5% level). [4] for both labels individually: MYO accuracy rises from 63.3% to 70.6% and LV accuracy rises from 81.9% to 89.9%. While this is for a single split, observe that using LVV as an auxiliary task effectively brought us closer to the range of having 50% annotated masks (second row in Table 6.4). Thus, auxiliary tasks, such as LVV prediction, which is related to the endocardial border segmentation, can be used to train models in a multi-task setting and leverage supervision present in typical clinical settings.

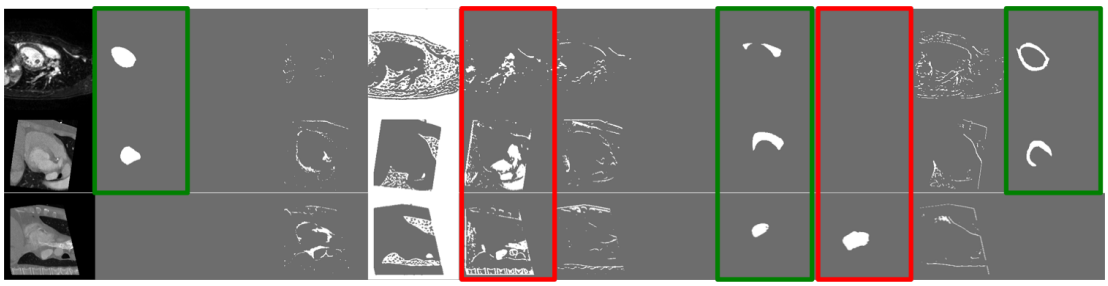


Figure 6.13: Example anatomical representations from one MR and two CT images respectively. Green boxes mark common spatial information captured in the same channels, whereas red boxes mark information present in one but not the other modalities.

6.5.3.3 Multimodal Learning

By design, our model separates the anatomy factor from the image modality factor. As a result, it can be trained using multimodal data, with the spatial factor capturing the common anatomical information and the non-spatial factor capturing the intensity information unique to each image's particular modality. Here we evaluate our model using a multimodal MR and CT input to achieve segmentation and modality transformation.

Both these tasks rely on learning consistent anatomical representations across the two modalities. However, it is well known that MR and CT have different contrasts that accentuate different tissue properties and may also have different views. Thus, we would expect some channels of the anatomy factor to be used in CT but not in MRI whereas some to be used by both. This disentanglement of information captures both differences in tissue contrasts but also differences in view when parts of the anatomy are not visible in all slice positions of a 3D volume.

This is illustrated in Figure 6.13, which shows three example anatomical representations from one MR and two CT images, and specifically marks common anatomy factors that are captured in the same respective channels, and unique factors that are captured in different channels.

Multimodal Segmentation

We train SDNet using MR and CT data with the aim to improve learning of the anatomy factor from both MR and CT segmentation masks. In fact, we show below that when mixing data from MR and CT images, we improve segmentation compared to when using each modality

[4]The multi-task objective in fact benefits the Dice score (statistically significant at the 5% level)

separately. Since the aim is to specifically evaluate the effect of multimodal training in segmentation accuracy, unlabelled images are not considered here as part of the training process, and the models are trained with full supervision only.

In Table 6.5 we present the Dice score over held out MR and CT test sets, obtained when training a model with differing amounts of MR and CT data. Results for 12.5% of data correspond to images obtained from one subject. Training with both data leads to improvements in both individual MR and CT performances. This is the case even when we add 12.5% of CT on 100% of MR, and vice versa; this improves MR performance (from 75% to 76%, not statistically significant, although improvement becomes significant as more CT are added), but also CT performance (from 77% to 81%, statistically significant).

We also train using different mixtures of MR and CT data, but keeping the total amount of training data fixed. In the CT case, we observe that Dice ranges between 77% (at 100%) and 65% (at 12.5%). This shows that CT segmentation clearly benefits from training alongside MR, since when training on CT alone with 12.5%, the corresponding Dice is 23%. In the MR case, we observe that Dice ranges between 75% (at 100%) and 49% (at 12.5%). Here, the relative reduction is larger than in the CT case, however MR training at 12.5% also benefits from the CT data, since the Dice when training on 12.5% MR alone is 27%. Furthermore, the Dice score for the other proportions of the data is relatively stable with a range of 69% to 74% for CT, and a range of 67% to 75% for MR.

In both experimental setups, whether the total number of training data is fixed or not, having additional data even when coming from another modality helps. This can have implications for current or new datasets of a rare modality, which can be augmented with data from a more common modality.

Modality Transformation

Although our method is not specifically designed for modality transformations, when trained with multimodal data as input, we explore cross-modal transformations by mixing the disentangled factors. This mixing of factors is a special case of latent space arithmetic that we demonstrate concretely in Section 6.5.3.5. We combine different values of the modality factor with the same fixed anatomy factor to achieve representations of the anatomy corresponding to two different modalities.

MR train	CT train	MR test	CT test
100%	100%	78_5	80_1
100%	12.5%	76_3	56_6
12.5%	100%	39_7	81_1
12.5%	0%	27_{12}	-
0%	12.5%	-	23_7
100%	0%	75_3	-
87.5%	12.5%	74_5	65_6
75%	25%	75_2	69_3
62.5%	37.5%	72_2	69_2
50%	50%	68_5	73_3
37.5%	62.5%	67_4	73_4
25%	75%	67_6	74_3
12.5%	87.5%	49_7	73_6
0%	100%	-	77_4

Table 6.5: Dice score (%) on MM-WHS (LV, RV, MYO, LA, RA, PA, AO) data, when training with different mixtures of MR and CT data. Standard deviations are shown as subscripts.

To illustrate this we use the model trained with 100% of the MR and CT in the MM-WHS dataset and demonstrate transformations between the two modalities. In Figure 6.14 we synthesise CT images from MR (and MR from CT) by fusing a CT modality vector z with an anatomy s from an MR image (and vice versa). We can readily see how the transformed images capture intensity characteristics typical of the domain. Note however, that as a result of the properties of the anatomy factor and the decoder, synthetic images appear smooth (no texture) and may lack realism. The anatomy factor resembles a multi-label segmentation, since each channel is binary and corresponds to a particular image region, whereas the decoder combines the anatomy and modality factor using FiLM, which applies affine transformations on feature maps. The above make it challenging for the decoder to synthesise texture within an anatomical region. This is also demonstrated in very recent work [15] by generating images from segmentation masks, where it is shown that synthetic quality can be improved with alternative decoder architectures. However, the goal of our approach is not cross-modal synthesis, and we use reconstruction costs to learn disentangled representations and drive semi-supervised learning.

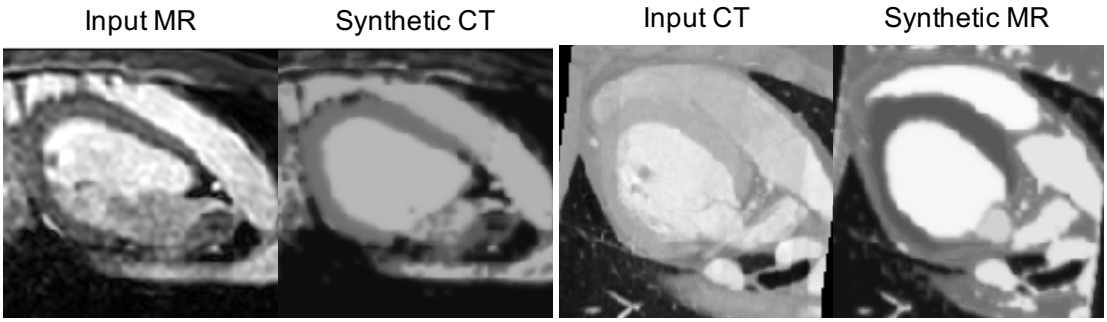


Figure 6.14: Modality transformation between MR and CT when a fixed anatomy is combined with a modality vector derived from each imaging modality. Specifically, let x_{MR}, x_{CT} be MR and CT images respectively. The left panel shows the original MR image x_{MR}, and a reconstruction of x_{MR} using the modality component derived from x_{CT}, i.e. $g(f_{anatomy}(x_{MR}), f_{modality}(x_{CT}, f_{anatomy}(x_{CT})))$. The right panel shows the original CT image x_{CT}, and its reconstruction using the modality of x_{MR}, i.e. $g(f_{anatomy}(x_{CT}), f_{modality}(x_{MR}, f_{anatomy}(x_{MR})))$.

6.5.3.4 Modality Type Estimation

Our premise is that the learnt modality factor z captures imaging specific information. We assess this in two different settings using multimodal MR and CT data and also cine-MR and CP-BOLD MR data.

Taking one of the trained models of Table 6.5 corresponding to a split with 100% MR (14 subjects of 2,837 images) and 100% CT images (14 subjects of 1,837 images)[5], we learn posthoc a logistic regression classifier (using the same training data) to predict the image modality (MR or CT) from the modality factor z. The learnt regressor is able to correctly classify the input images as CT or MR, on a held out test set (3 subjects of 420 images for MR and 3 subjects of 387 images for CT) 92% of the time. To find whether there is a single z dimension that captures best this binary semantic component (MR or CT) we repeat 8 independent experiments training 8 single input logistic regressors, one for each dimension of z. We find that z_5 obtains an accuracy of 82%, whereas the remaining dimensions vary from 42% to 66% accuracy. Thus, a single dimension (in this case z_5) captures most of the intensity differences between MR and CT which are global and affect all areas of the image.

[5]The results are based on a single split for ease of interpretation as between different splits we cannot relate the different z dimensions.

In a second complementary experiment we perform the same logistic regression classification to discriminate between cine-MR and CP-BOLD MR images (which are also cine, but contain additionally oxygen-level dependent contrast). Here, SDNet and the logistic regression model are trained using 95 cine-MR and 214 CP-BOLD images from 7 subjects, and evaluated on a test set of 27 and 31 images from 1 subject respectively. Unlike MR and CT which are easy to differentiate due to differences in signal intensities across the whole anatomy, BOLD and cine exhibit subtle spatially and temporally localised differences that are modulated by the amount of oxygenated blood present (the BOLD effect) and the cardiac cycle and these are most acute in the heart.[6] Even here the classifier can detect BOLD presence with 96% accuracy, when all dimensions of z are used. When each z dimension is used separately, accuracy ranges between 47% and 65%, and thus no single z dimension globally captures the presence (or lack) of BOLD contrast.

These findings are revealing and have considerable implications. First they show that our modality factor z does capture modality specific information which is obtained completely unsupervised, and depending on context and complexity of the imaging modality, a single z dimension may capture it almost completely (in the case of MR/CT).[7]

More importantly, it opens the question of how the spatial and modality factors interact to reproduce the output. We address these questions below using latent space arithmetic.

6.5.3.5 Latent Space Arithmetic

Herein we demonstrate the properties of the disentanglement by separately examining the effects of anatomical and modality factors on the synthetic images and how modifications of each alter the output. For these experiments we consider the model from Table 6.3, trained on ACDC using 100% of the labelled training images.

Arithmetic on the spatial factor s: We start with the spatial factor and in Figure 6.15 we alter the content of the spatial channels to qualitatively see how the decoder has learnt an association between the position of each channel and different signal intensities of the anatomical parts. In all these experiments the z factor remains the same. The first two images show the input and

[6] These subtle spatio-temporal differences can detect myocardial ischemia at rest as demonstrated in [10, 208].

[7] It is possible to detect the modality from the anatomy factor alone. If there are systematic differences between the modalities, this can be exploited by a classifier for detection. However, in this case the modality information is not actually contained in the anatomy factor.

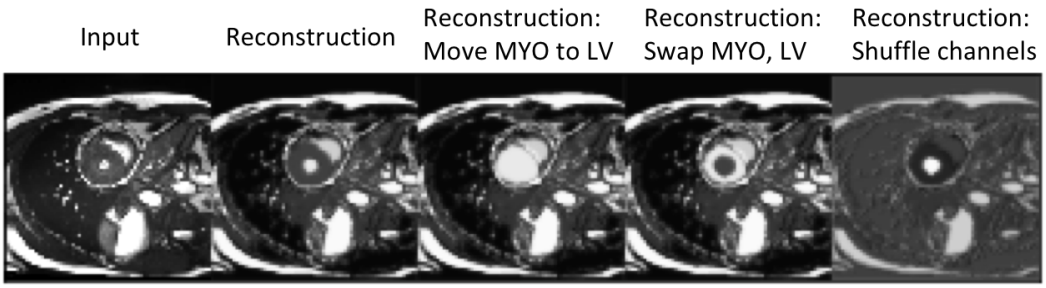


Figure 6.15: Reconstructions of an input image, when re-arranging the channels of the spatial representation. The images from left to right are: input, original reconstruction, reconstruction when moving the MYO to the LV channel, reconstruction when exchanging the content of the MYO and the LV channels, and finally a reconstruction obtained after a random permutation of the channels.

the original reconstruction. The third image is produced by adding the MYO spatial channel with the LV spatial channel and by nulling (zeroing) the MYO channel. We can see that the intensity of the myocardium is now the same as the intensity of the left ventricle. In the fourth image, we swap the channels of the MYO with the one of the LV, resulting in reverse intensities for the two substructures. Finally, the fifth image is produced by randomly shuffling the spatial channels.

Arithmetic on the modality factor z: Next, we examine the information captured in each dimension of the modality factor. Since the modality factor follows a Gaussian distribution, we can draw random samples or interpolate between samples in order to generate new images. In this analysis, an image x is firstly encoded to factors s and z. Since the prior over z is an 8-dimensional unit Normal distribution, 99.7% of its probability mass lies within three standard deviations of the mean. As a result, the probability space is almost fully covered by values in the range $[-3, 3]$. By interpolating each z-dimension between -3 and 3, and whilst keeping the values of the remaining dimensions and s fixed, we can decode synthetic images that will show the variability induced by every z-dimension.

To achieve this we consider a grid where each z dimension is considered over 7 fixed steps from -3 and 3. Each row of the grid corresponds to one of the 8 z dimensions, whereas a column a specific z-th value in the range $[-3, 3]$. This grid is visualised in Figure 6.16.

Mathematically described, for $i \in \{1, 2, \ldots, 8\}$ and $j \in \{1, 2, \ldots, 7\}$, an image in the i^{th} row

and j^{th} column of the grid is $g(s, z \odot v_i + (1 - v_i) \odot \delta_j)$, where $\odot$ denotes element-wise multiplication, v_i is a vector of length 8 with all entries 1 except for a 0 in the i^{th} position, and $\delta_j = -3 + 6(j - 1)$.

In order to assess the effect of z_i (the i^{th} dimension of z) on the intensities of the synthetic results, we calculate a correlation image and a difference image (for every row of results). The value of each pixel in the correlation image is calculated using the Pearson correlation coefficient between the interpolation values of a z_i and the intensity values of the synthetic images for this pixel:

$$\rho(z_i, y_{h,w}) = \frac{\sum_{j=1}^{7}(z_i^j - \bar{z}_i)(y_{h,w}^j - \bar{y}_{h,w})}{\sigma_{z_i}\sigma_{y_{h,w}}} \; \forall \; h, w \in H, W, \tag{6.13}$$

where h, w are the height and width position of a pixel, $\bar{z}_i$ is the mean value of z_i, $\bar{y}_{h,w}$ is the mean value of a pixel across the interpolated images, and σ_{z_i} and $\sigma_{y_{h,w}}$ denote the standard deviations. The difference image is calculated for each row by subtracting the image in the last column position on the grid ($\delta_j = 3$) with the first position on the grid ($\delta_j = -3$). [8]

In Figure 6.16, the correlation images show large positive or negative correlation between each z dimension and most pixels of the input image, demonstrating that z mostly captures global image characteristics. However, local correlations are also evident for example between z_1 and all pixels of the heart, between z_4 and the right ventricle and between z_5 and the myocardium. However, different magnitude changes are evident, as the difference image in the last column of Figure 6.16 shows. z_1 and z_4 seem to alter significantly the local contrast of the left and right ventricle, whereas small changes in the myocardium contrast are incurred by z_5. Some z dimensions, although correlated, do not seem to significantly affect the contrast of the image, thus indicating that a smaller number of dimensions would suffice for this dataset.

6.5.3.6 Factor Sizes

While throughout this section we used $C = 8$ and $n_z = 8$, it is worthwhile discussing the effects of these important hyperparameters as they determine the capacity of the model.

We have found through experiments that when $C > 8$ many channels are all zero. This ad-

[8] Note that in order to keep the correlation and the difference image in the same scale [-1, 1], we rescale the images from [-1, 1] to the [0,1], which does not have any effect on the results.

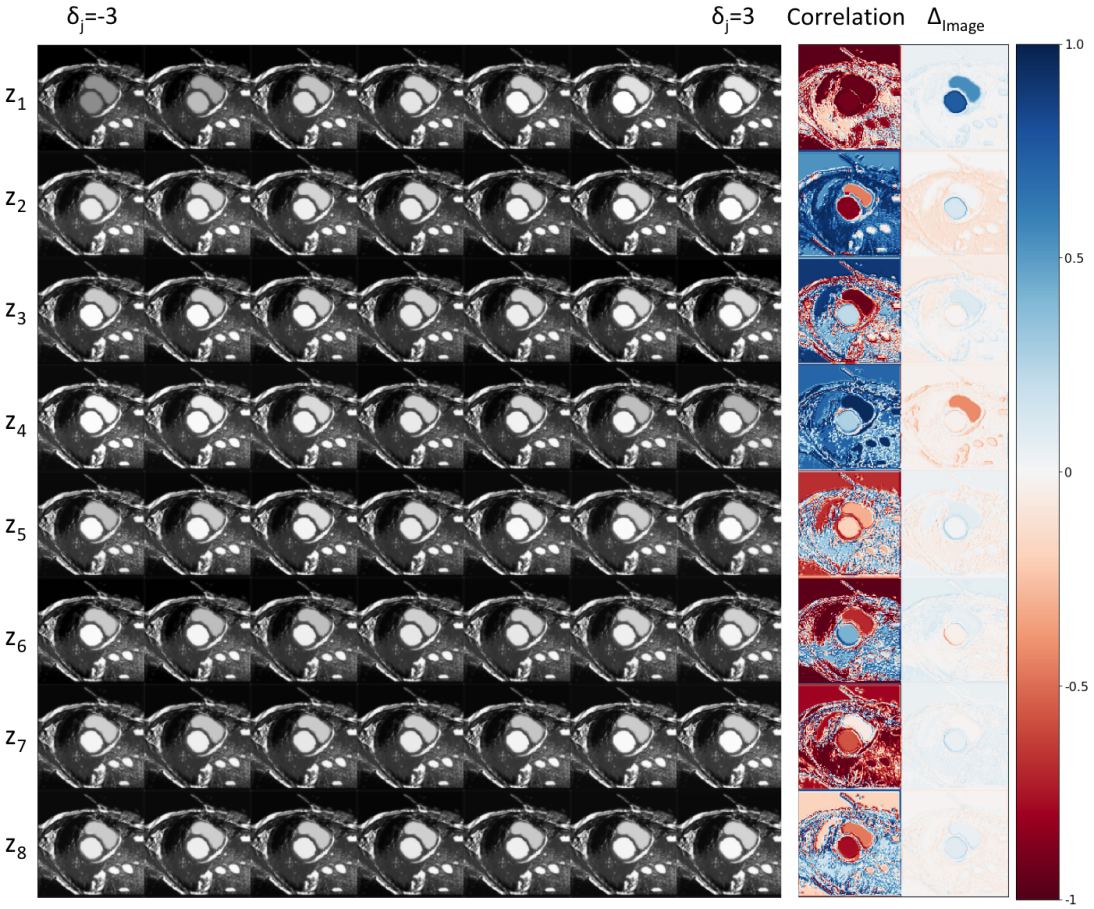


Figure 6.16: Reconstructions when interpolating between z vectors. Each row corresponds to images obtained by changing the values of a single z-dimension. The final two columns (correlation and Δ_{image}) indicate areas of the image mostly affected by this change in z.

ditional capacity is helpful when we use multimodal data, as for example in the MR/CT experiments, where $C = 16$. This allows to capture information common and unique across the two modalities in different s-channels see Figure 6.13). On the other hand, making C small ($C < 4$) we find that the model does not have enough capacity (for example an SDNet with $C = 4$ trained at 100% labels has Dice performance $68.1 \pm 8\%$, a drop compared to 84% when $C = 8$, that is also statistically significant at 5%).

We used $n_z = 8$ inspired by related literature [129]. Experiments with similar values of n_z maintain the segmentation performance, though this is decreased for high values of n_z. Specifically, an SDNet with 4, 32, and 128 dimensions trained at 100% labels has Dice $84 \pm 5\%$, $83 \pm 6\%$, and $82 \pm 6\%$, respectively. Compared to 84% when $n_z = 8$, the results for $n_z = 4$

and $n_z = 32$ are similar, but the result for $n_z = 128$ is worse (and also statistically significant at 5%), suggesting that the additional dimensions may negatively affect training and do not store extra information. To assess this we used the methodology in [209] to find the capacity of each z-dimension, which is also a measure of informativeness. This is calculated using the average variance per dimension, where a smaller variance indicates higher capacity. A variance near 1 (with a mean=0) would indicate that this dimension encodes a Normal distribution for any datapoint, and thus, according to [209], is uninformative and points to encoding the average of the distribution mode. Using this analysis, for $n_z = 128$ we observed that two z-dimensions each had variance of 0.88, while the remaining 126 had an average variance of 0.91. Repeating this analysis for $n_z = 32$, $n_z = 8$ and $n_z = 4$ we get the following results. For $n_z = 32$, two dimensions each has variances 0.78 and 0.79, while the remaining 30 dimensions have an average variance of 0.81. For $n_z = 8$, two z-dimensions each has variances 0.63 and 0.73, while the remaining 6 have an average variance of 0.75. Finally for $n_z = 4$, two dimensions have variances 0.62 and 0.65, and the average variance of the other two is 0.77, which are similar to the results of $n_z = 8$. This analysis shows that with smaller n_z, more informative content is captured in the individual z-dimensions, and thus a high n_z is redundant for this particular task.

6.6 Conclusion

We have presented two methods for disentangling images into spatial and (non-spatial) latent representations employing an image reconstruction cost, while promoting interpretable latent spaces. To the best of our knowledge this is the first work to investigate semantic spatial representation factorisation, in which one factor of the representation is inherently spatial and thus well suited to spatial tasks.

We firstly presented SMDNet, a method that decomposes cardiac images into a myocardial segmentation and a vectorised latent representation of the residual anatomy and modality information. We demonstrated its applicability in semi-supervised myocardial segmentation. In the low-data regime ($\approx 1\%$ of labelled with respect to unlabelled data) it achieves remarkable results, showing the power of the proposed learned representation.

We have also presented SDNet, a method for disentangling cardiac images into a semantically meaningful spatial factor of the anatomy and a non-spatial factor encoding the modality information. Moreover, through the incorporation of a variational autoencoder, we can treat our

method as a generative model, which allows us to also efficiently model the intensity variability of medical data.

We demonstrated the utility of SDNet in a semi-supervised segmentation task, where we achieve high accuracy even when the amount of labelled images is substantially reduced. We also demonstrated that the semantics of our spatial representation mean it is suitable for secondary anatomically-based tasks, such as quantifying the left ventricular volume, which not only can be accurately predicted, but also improve the accuracy of the primary task in a multi-task training scenario. We also show that the factorisation of the model presented can be used in multimodal learning, where both anatomical and imaging information can be encoded to create synthetic MR and CT images, using even small fractions of CT and MR input images, respectively.

The methods of this chapter focused on jointly training with labelled and unlabelled images from a single modality (semi-supervised learning), as well as with images from a secondary modality. The latter can be considered as multimodal learning, although no information from one modality directly benefits the other. An intuitive extension would be to investigate whether disentangled latent representations can be used to combine (or fuse) multimodal information, such that learning from one modality can explicitly aid another. This is presented in the following Chapter 7.

Chapter 7
Multimodal and Disentangled Representation Learning

7.1 Introduction

Chapter 6 described SDNet, a methodology for disentangling medical images in anatomical and imaging representations, and demonstrated a potential in using disentangled representations in multimodal learning through training with data of different modalities. In this chapter, we aim to further improve segmentation by leveraging information from other modalities in a multimodal and disentangled representation of anatomy and modality factors (see Figure 7.1). On the basis of our prior work of Chapter 4, we not only train with multimodal data, but also explicitly fuse disentangled anatomical features that are shared for all modalities. Indeed, disentangled representations are suitable for multimodal learning, since they can address many challenges posed by multimodal data. These include differences in signal intensities, a lack of

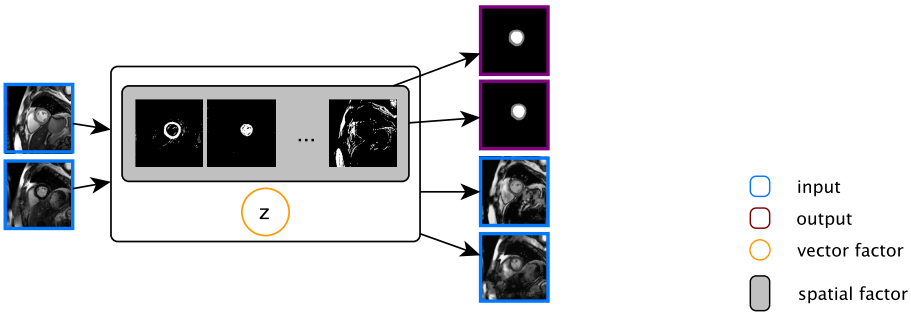


Figure 7.1: Multimodal and disentangled spatial and vector representations.

This chapter is based on:

- Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D., Dharmakumar, R., Tsaftaris, S.A., 2019. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE Transactions on Medical Imaging (under review).
- Chartsias, A., Papanastasiou, G., Wang, C., Stirrat, C., Semple, S., Newby, D., Dharmakumar, R., Tsaftaris, S.A., 2019. Multimodal Cardiac Segmentation Using Disentangled Representation Learning. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 128-137). Springer, Cham.

annotated data, as well as anatomical and temporal misalignments due to varying spatial resolutions or due to moving organs as in the case of dynamic imaging of the heart.

Multimodal learning, allows capturing information present in one modality (e.g. the anatomy) for use in another modality that has higher pathological contrast. As a motivating example, myocardial segmentation in LGE is challenging, since LGE mutes myocardial signal to accentuate signal originating from myocardial infarction. In fact, in clinical practice, analysis of LGE is typically combined with cine-MR [210].

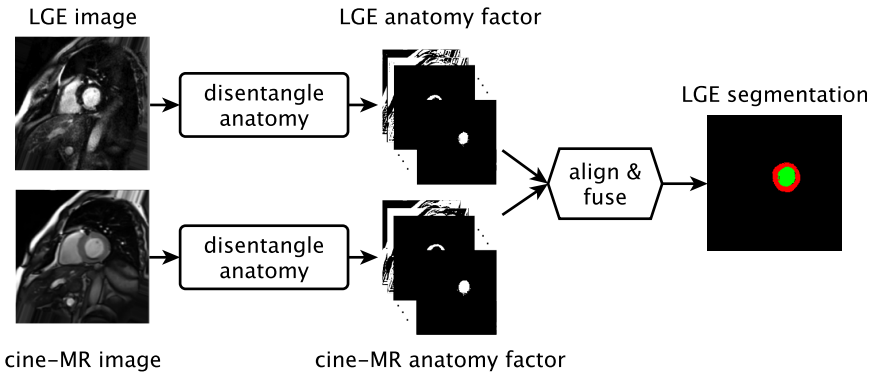
A naive way to propagate knowledge between modalities would be co-registration. This has been successful in the brain (see Section 3.5.3.3). But precise multimodal registration remains challenging, due to the need for modality independent metrics [63]. Critically, the brain remains static within an imaging session, whereas the heart is moving. Also, multimodal data are often inconsistent both in the number of images (different slices, cardiac phases, and perhaps more penalising resolution differences, e.g. slice thickness), as well as in the number of annotations. In addition, some sequences are static (LGE) and others dynamic (cine-MR). This necessitates solutions that alleviate misregistrations but also can pair input images.

7.1.1 Approach Overview

We propose a mechanism to represent data, that is suitable for learning how to propagate knowledge for segmentation. We learn both with and without annotations using a reconstruction objective. More excitingly, our approach co-registers data within an anatomical representation space, becoming thus robust to variations in imaging contrast. Our 2D approach, Disentangle Align and Fuse Network (DAFNet), see Figure 7.2, achieves the above by mapping multimodal images of the same subject into *disentangled* anatomy and modality factors.

Anatomy factors are represented as categorical feature maps. Each category corresponds to input pixels that are, ideally, spatially similar, and hence belong to the same anatomical part. This promotes semantic consistency and helps learn of spatial correspondences between anatomical parts from different modalities. Modality factors encode pixel intensities in a smooth multivariate Gaussian manifold as per the Variational Autoencoder (VAE) [44]. Anatomy factors are used to obtain segmentation masks, whereas their re-entanglement with the modality factors achieves image reconstruction.

A disentangled representation is encouraged by minimising the information capacity of each



and also in removing outliers.

Figure 7.2: DAFNet schematic in a LGE segmentation exemplar task using LGE and cine-MR inputs. Firstly, *disentangled* anatomy factors of the LGE and cine-MR image are extracted. Then, they are *aligned* (with a Spatial Transformer Network) and combined to a *fused* anatomy factor, used to infer the final segmentation mask. Our approach can use multi-input (multi-modal) data at training and inference. The latter is extremely useful when training with *zero annotations* for an input modality

factor respectively: thresholding the anatomy factors, prevents storing low intensity texture and imaging information, whereas the variational objective minimises the information in the modality factors [211]. Disentanglement is further influenced by the decoder design, either through inductive biases (see discussion in Section 7.3.4 and evaluation in Section 7.5.8), or through learning constraints (see cross-modal decoding of one anatomy in the modality of another in Section 7.3.4, similar to [25]).

However, a disentangled representation is not enough for multimodal learning. This ability comes from anatomy factors that are similar across modalities, and is achieved by weight sharing in the anatomy encoders, as well as by shared segmentation and decoder networks. These constraints implicitly create common anatomy semantics, which are essential when no labels, but only images, are available for one of the modalities. In this case we project all images to the common anatomical space, where a single segmentation network is trained with supervision only on the annotated modalities. When learning with multiple modalities, anatomy factors obtained from multimodal images are co-registered with a Spatial Transformer Network (STN) [188], fused with feature arithmetics, and also decoded in different modalities as defined by the modality factors. Finally, when input data are *not* paired (e.g. due to temporal or slice position differences) a new loss term in the cost function selects the most "informative"

multimodal pairs by comparing anatomy factors.

7.1.2 Contributions

Our contributions are the following:

1. We propose a 2D method for learning disentangled representations of anatomy and modality factors in multimodal medical images for segmentation.
2. We demonstrate the importance of semantic anatomy factors, that is achieved through the model design, since they allow learning registration and fusion operators.
3. We propose a loss term in the cost function that learns to select the most informative multimodal pairs.
4. We demonstrate our method's robustness over other approaches with extensive experiments on several datasets, in cardiac MRI and abdominal segmentation.
5. We show that our model works both on unimodal and multimodal inference, and that it outperforms other variants when trained with different amounts of annotations (semi-supervised) or zero annotations for one of the modalities.
6. We discuss different decoder designs using Feature-wise Linear Modulation (FiLM) [14] and Spatially-Adaptive (De)Normalization (SPADE) [15] respectively, and evaluate disentanglement by estimating the dependence between the anatomy and modality factor with distance correlation.

The remainder of this chapter is organised as follows. Section 7.2 presents recent related work on multimodal image analysis. Section 7.3 details the proposed model. Section 7.4 describes the data and benchmarks used. Section 7.5 presents the experimental results, and finally, Section 7.6 concludes with a discussion of the method.

7.2 Related Work

Multimodal machine learning is an active research area that involves learning with diverse sources of information. We consider multimodal learning as combining information of different images, present at training and/or inference time.

We review work on disentangled representations, the main focus of our method, and prior art on multimodal medical imaging for segmentation. We highlight though that currently no work exists that is able to simultaneously achieve multimodal fusion from unregistered data for image segmentation, be robust to the number of training annotations, and be applied to single or multimodal inference. These are made possible by the careful design of disentangled and semantic anatomical representations.

7.2.1 Disentangled Representation Learning

Our approach leverages learning disentangled anatomy and modality factors. Many approaches in computer vision [25] and in medical image analysis have been proposed for semi-supervised segmentation (SDNet in Chapter 6), multi-task learning [138], lung nodule synthesis [142], and registration [137]. Disentangling multimodal images has also been used for domain adaptation [145], although without applying information fusion.

As also discussed in Chapter 6, for anatomical features to be useful in clinical tasks, they are required to be semantic and quantifiable. This is not guaranteed in disentanglement techniques used for style transfer [25], or recent medical segmentation works [145] that do not impose restrictions on the content features. Differently from others, we disentangle quantifiable anatomical features, such that they are useful for segmentation, whereas interpretability is promoted with explicit design constraints (Section 7.3.1), which in addition enable registration and simple fusion operators.

7.2.2 Multimodal Learning

Multimodal learning is challenging in the presence of misaligned images. As results of Chapter 4 showed, a STN can be introduced in the learning process for performing affine transformations, in order to allow information fusion. Alternatively, a shared representation with both modalities can be learnt with encoder-decoder architectures. In particular, weight sharing of the layer closest to this representation yields the most effective results according to [189]. In cardiac analysis, most commonly contour models directly deform initialised segmentation masks [190,191]. Our method differs in that it does not require segmentation masks at inference time, and is the first to jointly learn suitable representation, co-registration, and information fusion for segmentation *but* in a semi-supervised setting.

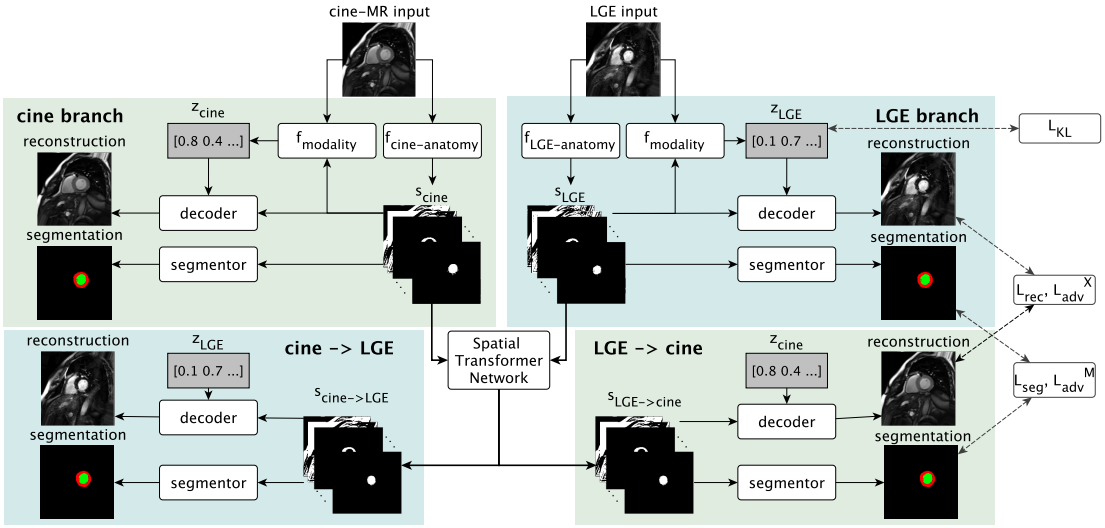


Figure 7.3: DAFNet training schematic with cine-MR and LGE input images. Each input is disentangled into anatomical and modality factors. With a STN the deformation branches ($cine \rightarrow LGE$, $LGE \rightarrow cine$) enable cross-modal synthesis and segmentation by deforming the anatomy factors s_{cine} and s_{LGE}. Losses are indicated on the right and are symmetrically applied to the cine-MR branch outputs on the left. L^z_{rec} is not shown. See text for definitions.

7.3 Proposed Approach

Here, we describe DAFNet, a multi-component 2D model for multimodal and semi-supervised learning that is robust to input misalignments. Inference consists of three stages. Firstly, encoders map images to anatomy and modality factors, then anatomy factors are spatially aligned and fused, and finally the fused factor produces segmentations.

Training is different with all involved costs and components illustrated in Figure 7.3. Input images are encoded into anatomy and modality factors (Section 7.3.1). Then, anatomy factors are aligned with a STN (Section 7.3.2), and also participate in segmentation losses (Section 7.3.3). Training further employs image reconstruction (Section 7.3.4) and modality reconstruction losses (Section 7.3.5). Finally a multimodal pairing loss allows to dynamically learn how to pair input image sources (Section 7.3.6). Below we detail the individual components, as well as the employed cost functions.

7.3.1 Encoding

Given modality i with samples $x_i \in X_i$, where $X_i \subset \mathbb{R}^{H \times W}$ is the set of images, and H and W are the height and width respectively, the encoding process achieves a disentanglement of anatomy and modality factors. Anatomy factors s_i are tensors produced by encoders dedicated to each modality i: $s_i = f_{anatomy}(x_i|\theta_i)$, where θ_i are the encoder parameters. The encoder architecture is modelled after the U-Net [13] and is shown in Figure 7.4. To reduce model parameters, and encourage a common anatomical representation among the multimodal data, we employ weight sharing in the decoder of each U-Net. Thus, the parameters θ_i are split into the unique parameters ϕ_i of the encoding path, and the shared parameters ψ of the decoding path: $s_i = f_{anatomy}(x_i|\phi_i, \psi)$.

An anatomy factor is represented as a binary tensor and thus cannot store different pixel intensities of an image as continuous values and this promotes the factorisation process. Furthermore, every pixel can be active at exactly one channel and this enforces a particular image region to appear in a single channel. More formally, $s_i \in \{0,1\}^{H \times W \times C}$, s.t. $\sum_{c=1}^{C} s_i^{h,w,c} = 1$ $\forall h \in \{1,\ldots,H\}, w \in \{1,\ldots,W\}$. Two anatomy factors produced by a cine-MR and an LGE image can be seen in Figure 7.5.

Anatomy factors are spatially aligned (Section 7.3.2), such that they can be fused and segmented, whereas in conjunction with the modality factor are decoded to synthetic images.

Divergence loss L_{KL}: The modality factors $z_i \in Z := \mathbb{R}^{n_z}$ are vectors produced by a single stochastic encoder, that, given an image sample x_i and its anatomy factor s_i, learns a probability distribution $q(z_i|x_i, s_i)$. In order to encourage a smooth space and minimise the encoded information [211], this posterior distribution is optimised to follow a multivariate Gaussian prior, $p(z_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, by minimising the $KL-$divergence with the re-parameterisation trick [44]:

$$L_{KL}(f_{modality}, f_{anatomy}) = \mathop{\mathbb{E}}_{x_i} \left[KL(q(z_i|x_i, s_i) \| p(z_i)) \right]. \tag{7.1}$$

The modality encoder is shown in Figure 7.4 and predicts the mean and standard deviation of a Gaussian that are used to draw the random sample vector z_i.

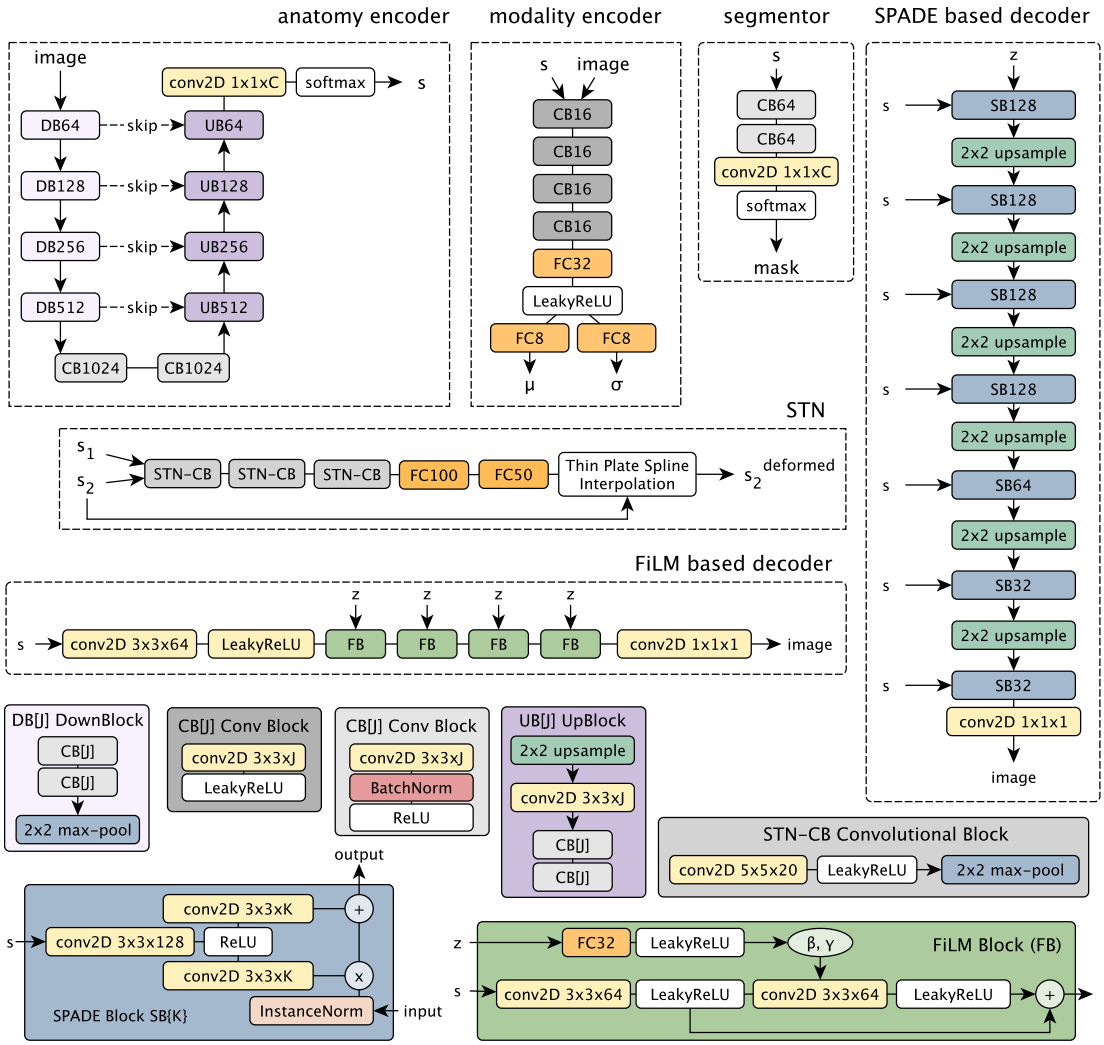


Figure 7.4: Architecture diagrams of the individual DAFNet components: the anatomy encoder extracts anatomy factors; the modality encoder extracts parameters μ, σ of a Gaussian distribution, and the modality factor is a sample from this distribution; the segmentation network produces a mask given an anatomy factor; a Spatial Transformer Network receives two anatomy factors and produces the 2D co-ordinates of 25 control points, used for interpolation; finally, two decoder architecture based on FiLM [14] and SPADE [15] decode anatomy and modality factors to images.

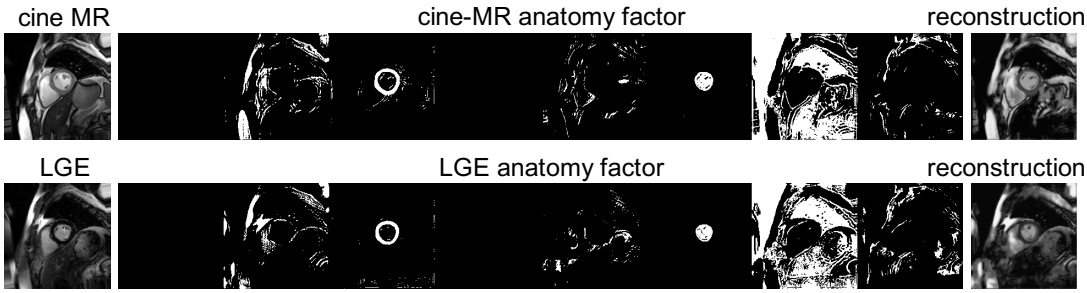


Figure 7.5: Anatomy factors from a cine-MR and a LGE. Observe how the same anatomical regions appear in the same channels.

7.3.2 Alignment and Fusion of the Anatomy factors

Following factor encoding, two anatomy factors s_i and s_j of modalities i and j respectively, are aligned using non-linear registration. Given an initial grid of 5×5 control points, a STN (architecture in Figure 7.4) first predicts the grid's offsets. Then, Thin plate spline interpolates the surface passing through the control points to register s_j with s_i. The result of the alignment step, $s_i^{deformed} = stn(s_j, s_i)$, is a deformed anatomy factor corresponding to s_j, and vice versa ($s_j^{deformed}$ corresponds to s_i). We optimise the STN with gradients in image space (see decoding cost of Section 7.3.4), as well as with the segmentation cost of Section 7.3.3, since we aim to align segmentation masks.[1]

During inference, the deformed anatomies are combined, to produce a fused representation containing all unique and shared features, that are present in the constituent anatomy factors. Since they are spatially aligned, a pixel wise operation such as the pixel-wise max is able to preserve all encoded features. More formally, $s_i^{fused} = \max(s_i, s_j^{deformed})$ and $s_j^{fused} = \max(s_j, s_i^{deformed})$. One benefit of max-fusion is that it is invariant to the number of inputs, and is therefore directly applicable in cases with more than two modalities.

7.3.3 Segmentation

Given an anatomy factor s_i, a convolutional network (architecture in Figure 7.4) infers a corresponding segmentation mask $m_i = h(s_i)$, s.t. $m_i \in M_i \coloneqq \{0, 1\}^{H \times W \times V}$, where M_i is the set of masks of modality i and V the number of classes. The segmentation network is common for

[1] We avoid direct comparison of s_i and $s_j^{deformed}$, since they are binary and thus different small deformations might generate the same error.

all modalities, is also applied to the deformed and fused anatomies, and is optimised as follows.

Supervised loss L_{sup}: Given a set of images paired with masks (x_i, m_i), a supervised cost is defined as a weighted sum of the differentiable Dice loss and Cross Entropy (CE):

$$L_{sup}(f_{anatomy}, h, stn) = \mathop{\mathbb{E}}_{x_i, m_i} \left[(1 - Dice(h(s_i), m_i)) + \alpha CE(h(s_i), m_i) \right], \tag{7.2}$$

where α control the balance between the losses and is set to $\alpha = 0.1$. The cross entropy and differentiable Dice are respectively defined as:

$$CE(h(s_i), m_i) = -\sum_{v} (m_{h,w,v} \log(p_{h,w,c})), \tag{7.3}$$

$$Dice(h(s_i), m_i) = 2 \times \left[\frac{\sum_{h,w,c} (h(s_{i_{h,w,c}}) \times m_{i_{h,w,c}})}{\sum_{h,w,c} (h(s_{i_{h,w,c}}) + m_{i_{h,w,c}})} \right], \tag{7.4}$$

where h, w, and c refer to the height, width and channel, and $p_{h,w,v}$ is the probability for a pixel belonging to class v.

Adversarial loss L^M_{adv}: An unsupervised segmentation cost is defined with a mask discriminator D_M, modelled after LS-GAN [50]. The adversarial objective given real masks sampled from all modalities $m \sim M_i, i \in \{1, 2, \dots n\}$ is:

$$L^M_{adv}(f_{modality}, h, stn) = \mathop{\mathbb{E}}_{x_i, m} \left[D_M(h(s_i))^2 + (D_M(m) - 1)^2 \right], \tag{7.5}$$

where the discriminator is adversarially trained against the segmentation network. The discriminator's architecture consists of 4 convolutional layers followed by LeakyReLU and a final single neuron layer, and uses Spectral Normalisation [53] to stabilise training. In both segmentation costs, the anatomy factors s_i come from either the input images directly, or are the result of the alignment step of a secondary j modality: $s_i \in \{f_{anatomy}(x_i|\theta_i), s_j^{deformed}\}, j \neq i$. In the latter case, the gradients produced by the segmentation cost are back-propagated to the STN module to learn its parameters.[2]

Training with segmentation losses helps learn better anatomy factors that separate the anatomies of interest in respective channels (see myocardium and left ventricle in Figure 7.5). If supervision is not available for modality i, training is performed with the adversarial loss L^M_{adv} and

[2]We omit the use of s_j^{fused} as input to the segmentation network to avoid backpropagating gradients both to the STN and the j^{th} anatomy encoder, which might result in the STN not achieving a good convergence.

with supervision of other modalities. This enables unsupervised segmentation of modality i, since factors are shared and common.

7.3.4 Decoding

The anatomy factors are further decoded into an output image of a style dictated by a modality factor z_i: $y_i = g(s_i, z_i)$. This can be performed with different decoders, which indirectly influence the type of disentanglement, or in other words the type of information captured by the anatomical and modality factors. We investigate two decoder architectures based on FiLM [14] and SPADE [15].

The input of the FiLM-based decoder (Figure 7.4) is the anatomy factors, which, after a series of convolutions, are conditioned by z samples. These are used to predict a scale and an offset parameter $\gamma \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$, which modulate each intermediate feature map $F \in \mathbb{R}^{H \times W \times C}$, where H, W and C are the height, width and number of channels respectively: $FiLM(F|\gamma, \beta) = F \odot \gamma + \beta$.

We also consider a SPADE-based decoder (Figure 7.4), which has been demonstrated to generate texture details on synthetic images given segmentation masks. The input to this decoder is a z sample, that is processed by a series of convolutional layers, conditioned by the anatomy factor, defining the output "shape". An Instance Normalisation layer with parameters μ and σ, is firstly applied to a feature map $F \in \mathbb{R}^{H \times W \times C}$, which is then modulated by tensors $\mathbf{\Gamma}$ and $\mathbf{B}$ (same size as F): $SPADE(F|\mathbf{\Gamma}, \mathbf{B}) = \mathbf{\Gamma} \odot \frac{F-\mu}{\sigma} + \mathbf{B}$.

Reconstruction cost L_{rec} and $L_{adv}^{X,i}$: The decoders are trained to reconstruct the input with the following loss:

$$L_{rec}(f_{anatomy}, f_{modality}, g, stn) = \mathop{\mathbb{E}}_{x_i} \left[\|x_i - g(s_i, z_i)\|_1 \right], \tag{7.6}$$

where $z_i = f_{modality}(x_i, s_i)$. In addition, synthesis of realistic images is encouraged with an adversarial loss of an image discriminator $D_{X,i}$ for each modality i (same architecture as D_M):

$$L_{adv}^{X}(f_{anatomy}, f_{modality}, g, stn) = \mathop{\mathbb{E}}_{x_i} \left[D_{X,i}(g(s_i, z_i))^2 + (D_{X,i}(x_i) - 1)^2 \right]. \tag{7.7}$$

As in the segmentation case, s_i is an encoding of image x_i or a deformed encoding of another image x_j: $s_i \in \{f_{anatomy}(x_i|\theta_i), s_j^{deformed}\}$. When $s_i = f_{anatomy}(x_i|\theta_i), j \neq i$, the model

acts as an auto-encoder. This is critical to allow the use of non-annotated images and enable semi-supervised learning. In the case, where $s_i = s_j^{deformed}$, the backpropagated gradients are used to train the STN module and also aid the factorisation process process (since the "style" of the output image $g(s_j^{deformed}, z_i)$, as specified by z_i, corresponds to modality i and not j). The decoder also participates in the following loss that promotes disentanglement.

7.3.5 Reconstruction of the Modality Factor Loss L_{rec}^z

In order to encourage disentanglement and also avoid posterior collapse of the modality factor, we reconstruct the modality factor of a synthetic image. This prevents the decoder from ignoring the z-factors and only use the anatomy factors. We minimise the reconstruction of the modality factor:

$$L_{rec}^z(f_{anatomy}, f_{modality}, g) = \mathop{\mathbb{E}}_{z,x} \left[\|z - f_{modality}(y, f_{anatomy}(y))\|_1 \right], \tag{7.8}$$

where z is a random sample from a unit Gaussian and y is the synthetic image produced by this z sample. Encouraging the use of modality factors by the decoder is further achieved by cross-reconstructing a deformed anatomy in a modality dictated by the corresponding z-factor.

7.3.6 Non-expert Pairing

Better multimodal fusion and STN registration will be achieved by multimodal image pairs $\{x_i, x_j\}$ that are more similar in terms of their spatial and temporal positions. In cases where the multimodal images are not expertly paired, DAFNet can automatically measure anatomical similarities with an optional cost, that directly compares the anatomy factors, and "selects" only the most informative image pairs.

During training, and given an image x_i and a set of K candidate images from modality j: $\{x_j^1, x_j^2, \ldots, x_j^K\}$, the multimodal segmentation and reconstruction losses for a sample x_i are weighted accordingly by K weights, s.t. $\sum_{k=1}^K w_k = 1$. Due to the semantics of the anatomy factors, and the fact that they are categorical, we can be directly evaluate their overlap in terms of the Dice score. The Dice for each pair, becomes the input to a small neural network υ of two fully connected layers that outputs the weights, and is similar to the temperature scaling technique proposed for calibrating classification outputs [212]. The segmentation and recon-

struction costs given network υ are the following:

$$L_{sup}(f_{anatomy}, stn, \upsilon) = \sum_{k=1}^{K} w_k L_{sup}(m_i, m_k), \tag{7.9}$$

$$L_{rec}(f_{anatomy}, f_{modality}, g, stn, \upsilon) = \sum_{k=1}^{K} w_k L_{rec}(x_i, y_k), \tag{7.10}$$

where $m_k = h(f_{anatomy}(x_j^k))$, and $y_k = g(s_j^{deformed,k}), z_i)$. By weighting the loss functions, the STN module does not need to learn deformations for all pairs, nor does it need to match slices with different anatomical content. At inference time, the most accurate segmentation is produced from the weighted sum of the fusion with different slices $s_j^{fused,1}, s_j^{fused,2}, \ldots s_j^{fused,k}$:

$$m_i = w_1 h(s_j^{fused,1}) + w_2 h(s_j^{fused,2}) + \ldots + w_k h(s_j^{fused,k}). \tag{7.11}$$

This optional weighting of the cost function is only used in unpaired data, and as shown in experiment 7.5.3 converges to the same result as manual pairing.

7.4 Experimental Setup

7.4.1 Training Details

The model is trained with a multi-component loss function, $L = 0.1 \cdot L_{KL} + 10 \cdot L_{sup} + L_{adv}^M + L_{rec} + L_{adv}^X + L_{rec}^z$. The weights of the individual loss components are selected experimentally, such that the errors are in the same value range. Nevertheless, we select a higher weight on L_{sup} to encourage separation of segmentation classes, since segmentation is a challenging task. Furthermore, a reduced L_{KL} weight prevents posterior collapse, in which the z factor is ignored by the decoder; however, an even lower L_{KL}, would not promote a Gaussian prior approximation, leading to a non-smooth intensity manifold. Number of s channels and z dimensions are set to $C = 8$ and $n_z = 8$ respectively, as in Chapter 6.

The code is written in Keras [202] and is available at `https://github.com/agis85/multimodal_segmentation`. We train with Adam (learning rate of 10^{-4}), and evaluate using Stochastic Weight Averaging [213] to reliably compare between different methods. Quantitative evaluation is performed on 3-fold cross-validation, where the training, validation and test sets correspond to the 70%, 15% and 15% of the data volumes.

7.4.2 Data

Experiments use multimodal datasets of a source and a target modality, rescaled to $[-1, 1]$.

1. For LGE segmentation, we use cine-MR and LGE data of 28 patients [2], acquired at Edinburgh Royal Infirmary (ERI) that are described in Section 2.6.2.5.

2. To evaluate robustness on different medical data, we use abdominal T1-dual inphase and T2-SPIR data from CHAOS dataset, described in Section 2.6.3.1. We resample to an x-y spacing of 1.89mm, and crop to 192×192 pixels. In total, there are 1594 images.

3. Finally, we evaluate BOLD segmentation with a dataset (shorthand BOLD) of cine-MR and CP-BOLD images of 10 canines, described in Section 2.6.2.4.

7.4.3 Baseline and Benchmark Methods

We consider the following baselines, which assume source masks being available at inference time. Their performance directly depends on these source masks. If predicted masks were used e.g. the result of a U-Net, an additional confounder would be introduced. Thus, we report numbers with ground truth masks for a bias-free estimate, which albeit is elevated.

1. A lower bound computes the Dice score between real masks of two modalities, and is also a measure of misalignment of the multimodal data. This is referred to as **copy**, and can be used for segmenting a target modality without annotations from the target modality.

2. This lower bound can be improved after registering the multimodal images and applying the registration field to the source masks. The deformation field is calculated by affine registration using mutual information, followed by symmetric diffeomorphic using cross-correlation [214]. This is referred to as **register**, and can also be used without annotations of the target modality. "Copy" and "register" are common in clinical evaluation.

3. Finally, as non-deep learning method we implemented a version of a non-coupled **active contour** model akin to the one in [191]. We initialised the contour using the "copy" above. For each dataset, via a grid search, we found optimal contour length, smoothness, and stepping hyperparameters as: for ERI [0.5, 0.15, 0.7], BOLD [0.01, 0.15, 0.7] and CHAOS [0.5, 0.15, 0.7], respectively.

We also consider the following deep learning benchmarks.

1. As a supervised benchmark, we train a UNet on annotated data of the target modality, and refer to it as **UNet-single**. We further re-train a UNet on mixed training data of all modalities to evaluate its capability of concurrently handling multimodal data, and refer to it as **UNet-multi**.

2. We train SDNet [32] with full or semi supervision on data of the target modality, and refer to it as **SDNet-single**. We also train SDNet by mixing multimodal data, as demonstrated in [32], and refer to it as **SDNet-multi**.

3. We get two final benchmarks by training Multimodal UNsupervised Image-to-image Translation (MUNIT) [25] for image translation. The first uses MUNIT to translate images from source to target modality [97], and the second translates multimodal images to a domain invariant space [145]. In both cases, segmentation is performed *post-hoc* with a UNet on the combined data. We refer to these approaches as **Translation** and **DADR** respectively.

4. Finally, we implement **DualStream** [189], the most recent Deep Learning based method for handling multimodal data which does not require registered data.

7.5 Results and Discussion

Sections 7.5.1 and 7.5.2 present segmentation results, assuming a source modality that always contains annotations during training. The *source* modality is cine-MR for ERI and BOLD datasets, and T1 for CHAOS. The *target* modality is LGE, BOLD and T2 for ERI, BOLD and CHAOS, respectively. Unless explicitly specified, DAFNet uses a FiLM-based decoder, and we report test Dice of the fused anatomies. We evaluate the effects of: *input pairing* (Section 7.5.3); *registration* (Section 7.5.4); and a *SPADE-based* decoder (Section 7.5.7). Section 7.5.8 evaluates *disentanglement* of each decoder design. Where appropriate, bold font denotes the best (on average) method and an asterisk (*) denotes statistical significance of paired t-tests ($p < 0.05$ assessed via permutations) comparing with the second best (to avoid multiple comparisons).

Methods	Train	Test	Masks in test	*100% target annotations* ERI LGE	ERI cine	BOLD BOLD	BOLD cine	CHAOS T2	CHAOS T1
copy	–	multi	Yes	67_{06}	67_{06}	80_{01}	80_{01}	71_{10}	71_{10}
register	–	multi	Yes	68_{07}	67_{05}	81_{04}	84_{05}	70_{07}	73_{05}
AC	–	multi	Yes	66_{15}	66_{13}	68_{02}	72_{05}	65_{22}	65_{22}
UNet	single	single	No	78_{04}	85_{08}	$\mathbf{91}_{01}$	89_{01}	$\mathbf{85}_{17}$	86_{05}
SDNet	single	single	No	80_{03}	84_{09}	89_{03}	88_{04}	83_{16}	85_{08}
UNet	multi	single	No	81_{03}	83_{08}	89_{03}	88_{02}	$\mathbf{85}_{15}$	$\mathbf{88}_{03}$
SDNet	multi	single	No	80_{05}	$\mathbf{86}_{05}$	89_{02}	87_{03}	$\mathbf{85}_{11}$	$\mathbf{88}_{01}$
DualStream	multi	single	No	80_{06}	$\mathbf{86}_{09}$	89_{09}	88_{02}	$\mathbf{85}_{16}$	85_{09}
Translation	multi	single	No	79_{06}	84_{05}	83_{06}	88_{02}	83_{09}	87_{06}
DADR	multi	single	No	79_{05}	83_{06}	88_{04}	86_{02}	84_{16}	72_{22}
DAFNet	multi	single	No	$\mathbf{82}_{03}$	$\mathbf{86}_{02}$	88_{01}	$\mathbf{91}_{02}$	83_{17}	$\mathbf{88}_{01}$
DAFNet	multi	multi	No	$\mathbf{82}_{03}$	84_{02}	$\mathbf{91}_{01}$	$\mathbf{91}_{01}$	$\mathbf{85}^{*}_{05}$	87_{01}

Table 7.1: Segmentation results on three datasets when full (100%) annotations are available. For each dataset we show results on the target modality assuming the other one is the source (and vice versa).

7.5.1 Multimodal Segmentation: Full and Zero Supervision Setting

The prime contribution of our work is the ability to learn and infer in a multimodal setting.Thus, we first demonstrate that multiple inputs at training and inference time benefit segmentation. Tables 7.1 and 7.2 present test Dice scores on three datasets for DAFNet and the benchmarks of Section 7.4.3. Two setups are evaluated, assuming either that annotations are available for the target modality or not.

In the 100% case shown in Table 7.1, training with multiple inputs improves accuracy, even when multimodal data simply constitute an augmented dataset. When segmenting the target modality, the usage of multiple inputs at inference time by DAFNet, obtains similar Dice as other benchmarks, but considerably reduces the standard deviation, such as in the CHAOS case from 11% to 5%.

Methods	Train	Test	Masks in test	*0% target annotations* ERI LGE	ERI cine	BOLD BOLD	BOLD cine	CHAOS T2	CHAOS T1
copy	–	multi	Yes	67_{06}	67_{06}	80_{01}	80_{01}	71_{10}	71_{10}
register	–	multi	Yes	68_{07}	67_{05}	81_{04}	84_{05}	70_{07}	73_{05}
AC	–	multi	Yes	66_{15}	66_{13}	68_{02}	72_{05}	65_{22}	65_{22}
UNet	single	single	No	n/a	n/a	n/a	n/a	n/a	n/a
SDNet	single	single	No	n/a	n/a	n/a	n/a	n/a	n/a
UNet	multi	single	No	38_{23}	68_{12}	68_{23}	85_{05}	–	–
SDNet	multi	single	No	61_{18}	73_{07}	80_{03}	85_{03}	51_{09}	63_{13}
DualStream	multi	single	No	38_{23}	68_{12}	68_{23}	85_{05}	–	–
Translation	multi	single	No	37_{23}	61_{13}	61_{10}	74_{07}	–	45_{11}
DADR	multi	single	No	46_{19}	63_{13}	68_{11}	85_{01}	–	49_{17}
DAFNet	multi	single	No	72_{06}	$\mathbf{78}_{05}$	78_{02}	82_{03}	72_{12}	$\mathbf{74}_{06}$
DAFNet	multi	multi	No	$\mathbf{74}^{*}_{04}$	76_{04}	$\mathbf{85}^{*}_{03}$	$\mathbf{86}_{02}$	$\mathbf{74}^{*}_{03}$	71_{06}

Table 7.2: Segmentation results on three datasets when zero (0%) target modality annotations are available. For each dataset we show results on the target modality assuming the other one is the source (and vice versa). Single input, single output models cannot be trained with no annotations and are thus marked with n/a. Furthermore, we choose to omit results marked with $-$, since training of these methods did not converge.

In the 0% case shown in Table 7.2, the (learned) benchmark methods fail to produce accurate target segmentations for all datasets. As expected, models trained only on the source modality learn modality-specific features, and as such cannot generalise to the unseen target modality. DAFNet on the other hand, consistently maintains a better average and smaller variance by leveraging information from the source modality. This is due to the aligning of the multimodal representations in the anatomy space, which allows the shared segmentor trained with supervision on the source, to also segment the target modality with "zero" supervised examples.

We then exchange the source and target modality and report the cine-MR and T1 Dice by training new models where appropriate. The CP-BOLD sequence that creates the BOLD data is very similar to cine-MR, showing anatomy, but has elevated T2 contrast (the BOLD effect) [10]. In

addition, these data are acquired in controlled experiments in mechanically ventilated subjects, with the cine-MR and BOLD images acquired one after the other in the protocol. Thus, all methods perform well in the 0% case of cine-MR segmentation, with multiple inputs further improving DAFNet performance. On the contrary, segmenting LGE (and T2) is more difficult, and the LGE Dice is overall lower than the cine-MR one in the single-output DAFNet with the difference being bigger in the 0% annotations case. As a result, this hurts the multi-output cine-MR Dice. This is expected since the benefit of multimodal segmentation comes when one modality is easier to segment.[3] Therefore LGE benefits when considering cine-MR images, but the contrary would only be beneficial in cine-MR reconstruction problems, e.g. in the presence of motion artefacts [215].

7.5.2 Semi-supervised Segmentation

Here we evaluate the sensitivity of all methods on different amounts of ground truth annotations available during training. Table 7.3 presents the average (across all labels) cross-validation test set Dice score. Exemplar test results are shown in Figure 7.6. The number of images for both source and target modalities are fixed, but the amount of target annotations varies. Sampling the amount of annotations is performed on a subject-level, to avoid having a mixture of annotated and non-annotated images of the same subject in the training set. The DAFNet results correspond to using multiple inputs at inference time.

Average Dice for all methods is comparable when the number of annotations is high, although DAFNet achieves the lowest variance. With a reducing number of annotations, the performance of the competing methods also reduces with a simultaneous increase in the variance. DAFNet maintains good results and robustness to edge cases, as evidenced by the small variance achieved throughout all setups.

7.5.3 Effect of Pair Matching

The results of Sections 7.5.1 and 7.5.2 correspond to expertly paired multimodal inputs. Here, we evaluate the sensitivity of DAFNet on unpaired multimodal images, as well as the effect of the automated pairing cost proposed in Section 7.3.6.

[3] Indeed, cine-MR is designed to show anatomical information, whereas LGE to highlight infarcted myocardium.

Methods	*ERI: Target LGE*			*BOLD: Target BOLD*			*CHAOS: Target T2*		
	50%	25%	12.5%	50%	25%	12.5%	50%	25%	12.5%
copy	67_{06}	67_{06}	67_{06}	80_{01}	80_{01}	80_{01}	71_{10}	71_{10}	71_{10}
register	68_{07}	68_{07}	68_{07}	81_{04}	81_{04}	81_{04}	70_{07}	70_{07}	70_{07}
AC	66_{15}	66_{15}	66_{15}	68_{02}	68_{02}	68_{02}	65_{22}	65_{22}	65_{22}
UNet-single	76_{12}	66_{14}	51_{21}	79_{17}	59_{27}	49_{29}	80_{17}	76_{15}	72_{17}
SDNet-single	76_{04}	69_{09}	54_{18}	84_{03}	68_{17}	64_{14}	82_{14}	77_{16}	75_{14}
UNet-both	76_{08}	67_{11}	50_{19}	$\mathbf{87}_{03}$	75_{17}	72_{13}	$\mathbf{84}_{15}$	79_{16}	75_{16}
SDNet-both	76_{04}	73_{07}	64_{19}	86_{07}	85_{03}	80_{03}	$\mathbf{84}_{11}$	80_{13}	78_{09}
DualStream	76_{03}	61_{13}	44_{23}	86_{01}	58_{26}	49_{28}	81_{19}	78_{16}	75_{16}
Translation	75_{07}	67_{14}	62_{14}	84_{02}	79_{06}	47_{26}	81_{07}	75_{11}	70_{10}
DADR	77_{05}	66_{11}	57_{19}	$\mathbf{87}_{02}$	79_{01}	71_{15}	$\mathbf{84}_{11}$	77_{14}	74_{11}
DAFNet	$\mathbf{78}^{*}_{04}$	$\mathbf{76}^{*}_{05}$	$\mathbf{74}^{*}_{05}$	$\mathbf{87}_{01}$	$\mathbf{86}_{03}$	$\mathbf{85}^{*}_{03}$	$\mathbf{84}_{05}$	$\mathbf{82}_{03}$	$\mathbf{79}^{*}_{05}$

Table 7.3: Segmentation results of LGE, BOLD and T2, when training with a varying amount of annotations for ERI, BOLD, and CHAOS datasets respectively.

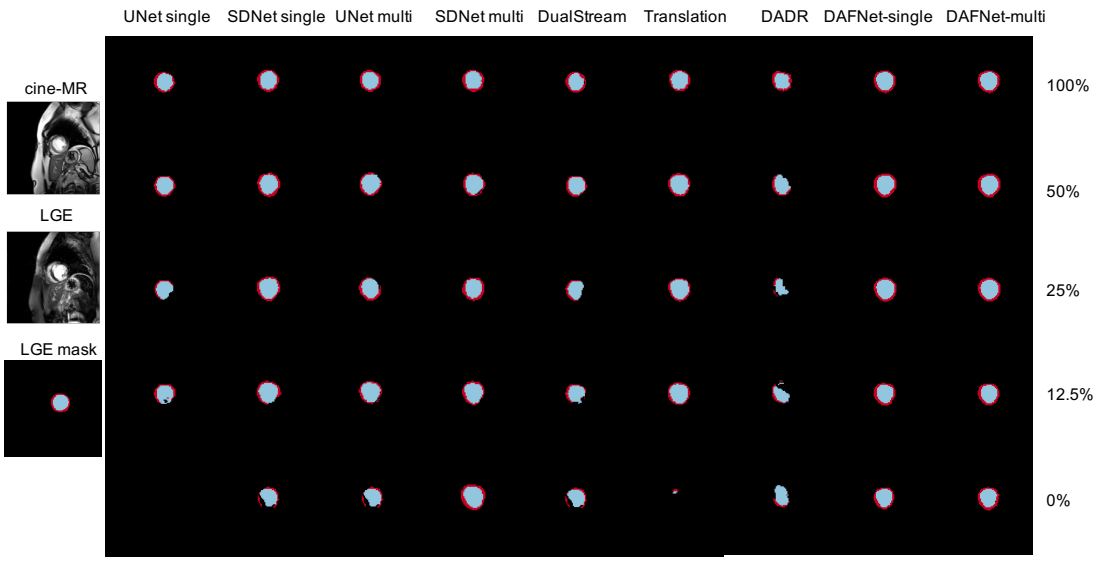


Figure 7.6: Panel of LGE segmentation examples from ERI dataset, obtained with different amount of LGE annotations.

We randomly shuffle the multimodal pairs by two positions, with the shuffled pairs differing up

Pair matching	copy	DAFNet 0%	DAFNet 100%
expert	67_{06}	74_{04}	82_{03}
automated	n/a	71_{06}	80_{03}
random	44_{16}	65_{08}	77_{06}

Table 7.4: LGE segmentation results when the multimodal images are not expertly paired.

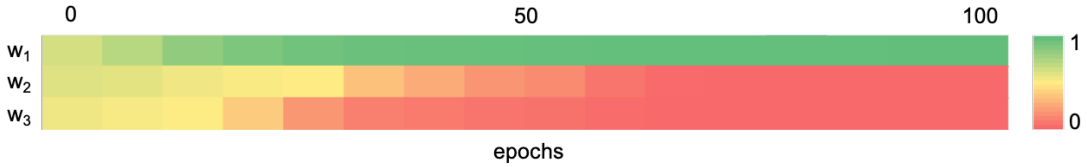


Figure 7.7: Evolution of weights w_j across epochs. Weights are used as a measure of similarity between each candidate multimodal pair. For more details see text.

to two spatial slices within a 3D volume.[4] We measure the LGE segmentation Dice score on ERI data when using 100% and 0% LGE annotations. We thus compare our automated method with expert pairing (upper bound) and a random shuffle (lower bound). Table 7.4 presents the results of copy method, as well as of DAFNet evaluated with both cine-MR and LGE inputs.

Shuffling the multimodal pairs decreases the copy performance considerably. In both cases automated matching of candidate pairs based on the semantics of the anatomy factors proves effective in ignoring distant slices (in the volume) with results very closely approaching the ones achieved by expert pairing. As described in Section 7.3.6, DAFNet weighs the contribution of each candidate slice to the fused representation. To show how appropriate weights are learnt, given an LGE image, we plot the evolution of three weights corresponding to three candidate cine-MR images across training epochs in Figure 7.7, where w_1 corresponds to the closest cine-MR image and w_3 to the most distant one. It can be seen that the weight w_1 converges to one early on in training, suggesting that the model is ignoring the more distant candidate images.

During inference, a "soft" segmentation mask is produced as a weighted sum between each weight with its corresponding mask. However, this converges to using the prediction of the "closest" pair, as evidenced by Figure 7.7.

[4]Similar results can be obtained by shuffling the different cardiac phases in the cine-MR temporal stack.

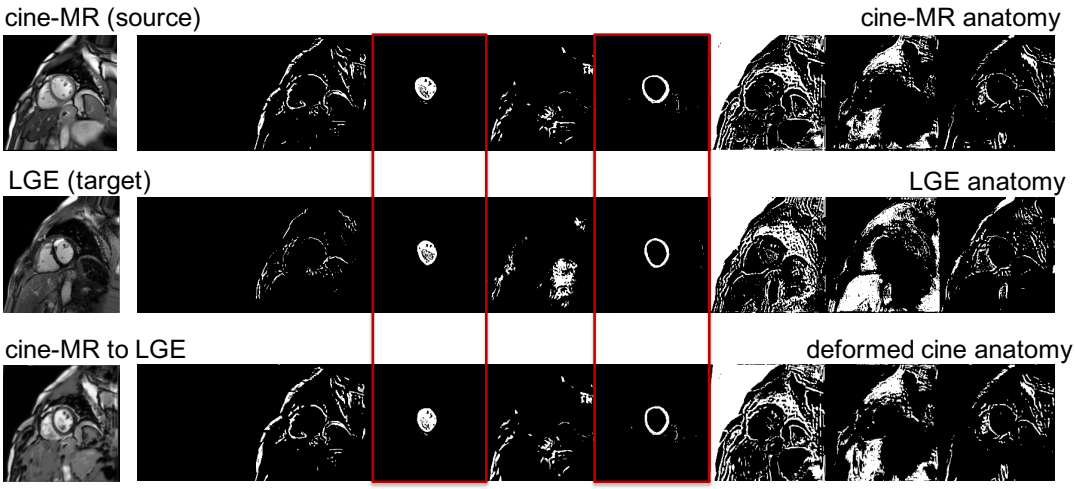


Figure 7.8: Example anatomy alignment. The source cine-MR anatomy (row 1) is deformed by the STN to match the target LGE one (row 2), resulting in the one of the last row. Red boxes mark channels of the areas of interest (left ventricle and myocardium).

7.5.4 Effect of STN

We assess the need for a registration module with an ablated model. We compare the accuracy of a fused segmentation that is obtained with and without the STN module. Two DAFNet models are compared, trained on ERI data with 100% and 0% LGE annotations. The mean Dice without the STN is measured to be $75 \pm 6\%$ and $71 \pm 6\%$ respectively. This is lower than the Dice of DAFNet with STN that is $82 \pm 3\%$ and $74 \pm 4\%$. Furthermore, in the 100% case the difference is statistically significant at the 1% level. Thus, clearly registration helps.

An example anatomy alignment is shown in Figure 7.8. Although not a perfect alignment of the images is required, the left ventricle and myocardium of the cine-MR have deformed to match the corresponding LGE (marked in red boxes).

7.5.5 Ablation Study on Cost Components

We assess the contribution of critical cost components in the fused segmentation on ERI with 100% and 0% LGE annotations. We evaluate the effect of adversarial training on masks, L^M_{adv}, and images L^X_{adv}, respectively, and the effect of modality factor reconstruction, L^z_{rec}.[5] Table 7.5

[5]We do not include ablations of the supervised segmentation, L_{sup}, the KL-divergence, L_{KL}, and the reconstruction cost, L_{rec}. We omitted L_{sup} because DAFNet is not fully unsupervised. Training without the L_{KL}

L_{adv}^{M}	L_{adv}^{X}	L_{rec}^{z}	DAFNet 0%	DAFNet 100%
—	✓	✓	72_{02}	80_{03}
✓	—	✓	65_{05}	71_{17}
✓	✓	—	71_{04}	81_{03}
✓	✓	✓	74_{04}	82_{03}

Table 7.5: Ablation study on the effect of individual cost components on LGE segmentation.

shows that best results are achieved with all cost components. Learning a data-driven reconstruction cost (via the image discriminator) contributes the most: by encouraging more accurate synthesis it helps to learn better anatomical representations.

7.5.6 Ablation on factor sizes C and n_z

In all experiments, factor sizes are set to C=8, and n_z=8. C is determined experimentally, such that there is enough capacity for all segmentation classes and background anatomy. A large C does not affect segmentation, and the redundant capacity is ignored, see "empty" channels of Figure 7.5. This is confirmed with ablated models with C=4 or C=16 trained with 100% annotations on ERI and CHAOS. The ERI model achieves $82\pm2\%$ for both setups, the same as when C=8. The CHAOS model achieves $74 \pm 12\%$ and $85 \pm 5\%$ for C=4 and C=16, respectively. The performance significantly drops when C=4, since there is not enough capacity.

Size n_z is determined according to our previous [32], and related work [25]. We experimented with n_z=4 and n_z=16 and 100% annotations on ERI and CHAOS. We find no effect on segmentation accuracy. However, n_z affects the information capacity, approximated by the average variance [209], of each z-dimension, where smaller variance implies higher informativeness. With $n_z = 16$, the lowest variance is 0.63, the first 8 dimensions have an average of 0.86 and the remaining 8 an average of 0.95. For n_z=8, the variance ranges between 0.47 and 0.80, and for n_z=4, between 0.43 and 0.60. Admittedly, lower n_z results in higher information content in each dimension, thus large n_z seems redundant in this setup.

and L_{rec} significantly change the model to one lacking a smooth modality space and the ability for cross-modal synthesis.

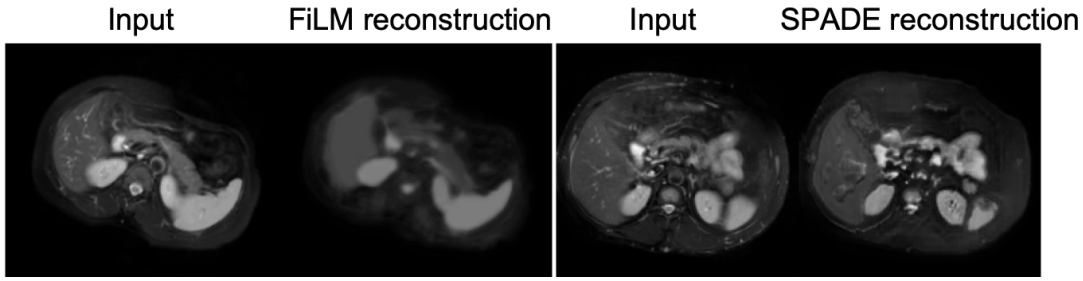


Figure 7.9: Reconstructions with two decoders. The FiLM synthetic image is more flat and lacks texture, in contrast to the SPADE synthetic image. Images taken from CHAOS dataset.

7.5.7 Effect of Decoder Design on Segmentation Accuracy

The modular design of DAFNet permits incorporation of components with different designs. We evaluate segmentation accuracy achieved by two decoder architectures: FiLM and SPADE. Specifically, we train a SPADE-based DAFNet on ERI and CHAOS and compare with the FiLM-based DAFNet for 100% and 0% annotations.

With 100% annotations, the SPADE-based DAFNet achieves $82 \pm 3\%$ and $85 \pm 5\%$ on ERI and CHAOS respectively, identical to the Dice achieved by FiLM. With 0% annotations, the SPADE-based DAFNet achieves $73 \pm 4\%$ and $75 \pm 7\%$, whereas FiLM-based results are $74 \pm 4\%$ and $74 \pm 3\%$ respectively on ERI and CHAOS.

We conclude that the regularising effect of the reconstruction process on extracting segmentations is similar in both decoder variants. However, different decoder designs influence the way the anatomy and modality factors interact to produce a synthetic image. We explore this next.

7.5.8 Evaluating Disentanglement

Even though FiLM and SPADE decoders do not result in evident differences in segmentation accuracy, they produce synthetic images of different quality (Figure 7.9). Since the anatomy factors contain flat regions, FiLM-based conditioning with scalar parameters tends to produce images with less texture details than SPADE-based conditioning.

Here, we aim to assess the information retained in the modality factors, and characterise the achieved disentanglement. This is a challenging problem not addressed in existing literature: all assume vector latent variables (e.g. BetaVAE score [44]). In DAFNet, and typically in

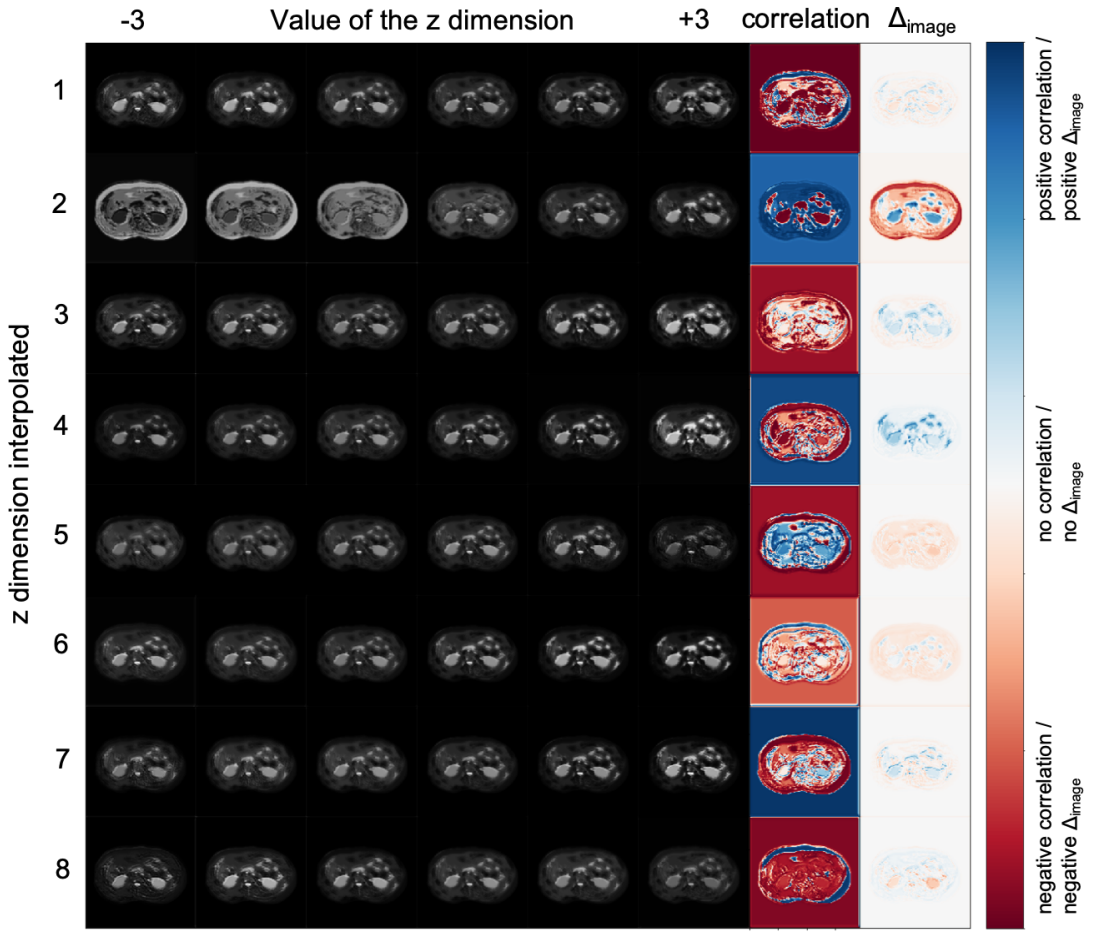


Figure 7.10: FiLM based reconstructions. Images per row correspond to interpolating a single z dimension. Last two columns (correlation, and difference image Δ_{image}), indicate regions mostly affected by each z dimension.

content/style disentanglement, the factors of variation are not of the same dimensionality, with the anatomy being spatial. For the experiments below, we use models trained on CHAOS with 100% T2 annotations to assess (dis)entanglement using classification tests, factor arithmetics, and a proposed metric of independence of random variables.

7.5.8.1 Modality Classification

On the premise that the common modality encoder correctly extracts modality features, a classifier should detect the modality type, given just the z-factor. We assess this hypothesis, by training a logistic regression classifier to predict whether different z-factors correspond to T1

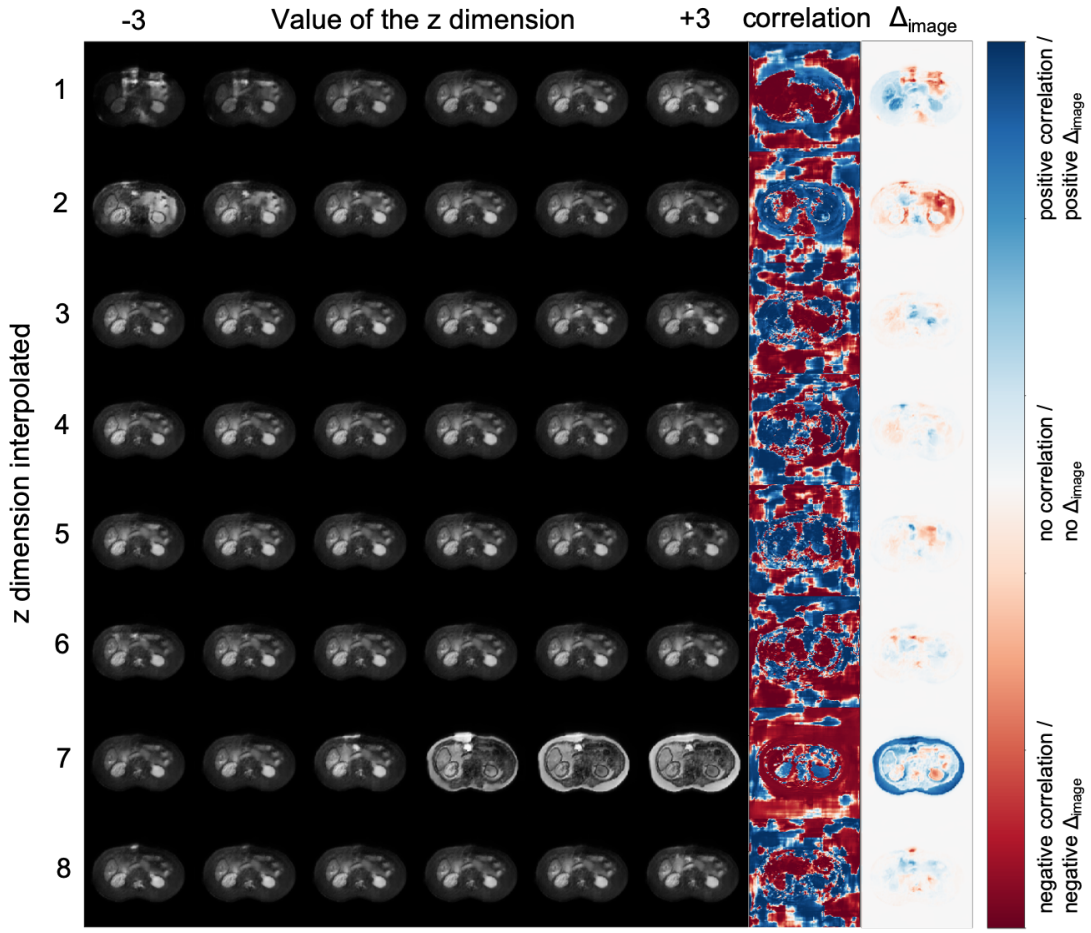


Figure 7.11: SPADE based reconstructions. Images per row correspond to interpolating a single z dimension. Last two columns (correlation, and difference image Δ_{image}), indicate regions mostly affected by each z dimension.

or T2 images. The classifier's accuracy is 99% and 97% for FiLM and SPADE, respectively, on a test set of three subjects.

We further evaluate whether specific dimensions in z capture the modality type by repeating the experiment, for each dimension. In the FiLM model, the 2nd dimension achieves 100% accuracy, whereas the rest vary between 54% and 64%. Similarly in the SPADE model, the 7th dimension achieves 97% accuracy vs. 42% and 63% of the others.

7.5.8.2 Modality Factor Arithmetics

We qualitatively examine the information retained in each dimension in vector z with latent space arithmetics. The likelihood of the modality factor approximates a Gaussian prior, and therefore interpolating in the range $[-3, 3]$ covers the probability space. Figures 7.10 and 7.11 shows synthetic images arranged in a grid; images of each row are produced by interpolating the values of a single dimension of z, with the remaining ones fixed. The final two columns highlight affected regions by calculating the per-pixel Pearson correlation, as well as the difference, Δ_{image}, between the synthetic images at extreme z values -3 and 3, respectively.

Both decoders have one z-dimension that has a global image effect (z_2 and z_7 respectively) and controls the modality type. This finding is inline with the classification results above. Furthermore, some dimensions of the FiLM decoder appear to be focused on specific anatomical regions, such as z_7 and z_8, which affect the contrast of the left and right kidneys. In contrast, the dimensions of the SPADE decoder produce more diffused correlation images. The same is observed on the difference images, where specific z-dimensions affect areas of the image not necessarily related to anatomical regions, such as z_1 and z_2. The latter is likely related to the SPADE architecture, which uses z as input to encode information on the image layout without a semantic correspondence to the anatomical layout of the anatomy factor. This helps with generating texture, but means that z-dimensions do not condition meaningful regions. Finally, in both FiLM and SPADE decoders, some z-dimensions do not have a significant effect on the contrast of any image regions, thus indicating that fewer dimensions could have been used.

7.5.8.3 Disentanglement Metric

We propose the use of distance correlation [216], as a metric of factor independence (and disentanglement), which is invariant to the input variable dimensionality, and can also detect linear and non-linear associations. While distance correlation has been used before for reducing data leakage [217], we use it here for measuring (dis)entanglement. Distance correlation is:

$$dCor(s, z) = \frac{dCov(s, z)}{\sqrt{dVar(s)dVar(z)}}, \tag{7.12}$$

where $dCov(s, z)$ is the distance covariance of s and z, and $dVar(.)$ is the distance variance respectively. Given n random samples s_k and z_k with $k \in [1, n]$, the distance covariance is the product of two distance matrices (one for each variable) averaged by n^2, where each

distance matrix $d(.)$ is double centred by subtracting the mean row, the mean column and the overall mean from each element: $dCov^2(s,z) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(s_i,s_j)d(z_i,z_j)$. The distance variance is then $dVar^2(s) = dCov^2(s,s)$, and $dVar^2(z) = dCov^2(z,z)$.

The distance correlation between s and z values from a FiLM-based model is $dCor = 0.55$, whereas the equivalent for a SPADE-based model is $dCor = 0.78$. This suggests that the anatomical and modality factors obtained by a FiLM decoder are more independent, and therefore the FiLM-based model is more disentangled. Although distance correlation cannot explicitly evaluate the type of information in each variable, this result can be explained intuitively by the decoder design. The SPADE decoder allows more flexibility to the z factors, and this is evident both in the synthetic images, which contain more texture, and also in the diffused correlation images of Figure 7.10, implying a higher anatomical correlation (and higher entanglement) between the z and s factors.

7.6 Conclusion

We have presented a method for multimodal learning, and specifically multimodal segmentation, that is robust to the requirement for registered and paired input images. This has been made possible by disentangling images into semantic anatomy factors, that are consistently represented across modalities, and modality factors that model the intensity variability of the multimodal inputs into a smooth latent space.

This chapter combined the findings of Chapters 4–6 for multimodal and disentangled spatial representations in a unified framework. We proposed DAFNet, which, to the best of our knowledge, is the first work that enables multimodal segmentation by aligning disentangled anatomical representations, and can be trained with few or zero annotations for one of the modalities.

We presented the benefit of multimodal (over unimodal) learning in cardiac and abdominal segmentation, where we achieve high accuracy and low variance through the fusion of anatomical information of different modalities. We further demonstrated robustness to misalignments in the multimodal data (achieved by a Spatial Transformer Network), and robustness to the quality of the multimodal pair matching (with an optional pair weighting), both made possible by comparing the semantic anatomy factors. Finally, we made a first step in evaluating the quality of the content/style disentanglement using the distance correlation, although limitations remain in the precise quantification of the type of information that is captured by each factor.

Chapter 8
Summary and Future Directions

This final chapter summarises the thesis contributions, discusses the significance of our work in Section 8.1, and presents some limitations and avenues for the future in Section 8.2.

8.1 Summary

This thesis considered deep learning methods for medical image analysis, specifically for the tasks of synthesis and segmentation. We proposed new methods that contribute to the medical imaging research through their ability to combine complementary information from multimodal images, as well as through their robustness to the number of annotations with semi-supervised learning. We investigated spatial representations, and demonstrated that they are suitable latent variables for learning cross-modal and multimodal correlations, as well as for representing factors of variation. Finally, we have contributed to the deep learning research by introducing disentangled spatial representations as a way of separating structural and appearance information from images. In brief we consider various image domains (modalities), and propose different methods of encoder-decoder architecture that address the problems of synthesis, segmentation, multimodal, and semi-supervised learning.

Chapter 4, explores multimodal synthesis by subsequently encoding and decoding images through intermediate spatial representations. This method proves that images, i.e. multi-channel feature maps, can be latent variables, and are suitable for synthesis problems of cardinalities one-to-one, many-to-one, or many-to-many. Robustness to the number of inputs makes the method applicable to scenarios with imperfect datasets, and feature fusion is important for leveraging complementary information. However, this method requires paired data for training.

Chapter 5 overcomes the data pairing problem by learning one-to-one mapping functions between image domains with the cycle consistency principle, and demonstrates the utility of synthetic images for data augmentation in auxiliary tasks. Our investigation shows that cycle consistency is problematic when the domains do not have similar information capacity.

Chapter 6 solves this information matching problem by introducing disentangled representations to encode the residual information of the lossy domain. Here, the mapping is bidirectional between images and anatomical semantic maps. Thus the residual information corresponds to appearance, i.e. pixel intensities of the medical images. In the proposed method, the anatomical map is part of the disentangled factors, thus allowing visual inspection of what a network learns. The presence of supervised segmentation, and unsupervised reconstruction losses make disentangled representation methods directly applicable to semi-supervised tasks, by taking advantage of the utility and semantics of the spatial content. This scheme however presumes single-domain images and segmentation labels, and cannot combine multimodal information.

Chapter 7, inspired by the multimodal synthesis method of Chapter 4, as well as by the findings of Chapter 6 on disentangled representations, uses multiple encoders to map multimodal images in a shared disentangled representation space. The common semantics of the anatomy factors enable information fusion. Also, image misalignments that are common in multimodal medical datasets, can be corrected in the spatial anatomy space, while the disentangled representation offers the ability for semi-supervised learning. In addition, we show that the learning costs and design biases also allow training in the absence of annotations for one modality.

The broader significance of our work is the disentanglement of medical image data into meaningful spatial and non-spatial factors. This intuitive factorisation does not require the specific network architecture choices used in this thesis, but rather is general in nature and thus could be applied in diverse medical image analysis tasks. Already experimental results of Chapter 6 have demonstrated the potential of combining imaging and non-imaging data, such as the ones available in electronic health records. Factorisation facilitates manipulations of the latent space and as such probing and interpreting the model. Such interpretability is considered key to advance the translation of advanced machine learning methods in healthcare.

8.2 Limitations and Future Directions

Our work has some limitations that inspire future directions. We can envision that extensions to 3D (in lieu of 2D), would further improve applicability of our approaches in several domains such as brain (which benefits from 3D view) and abdominal imaging. However, 3D models also present challenges. The number of model parameters significantly increases, since 3D convolutions are employed. Also, the effective size of datasets decreases, since each subject

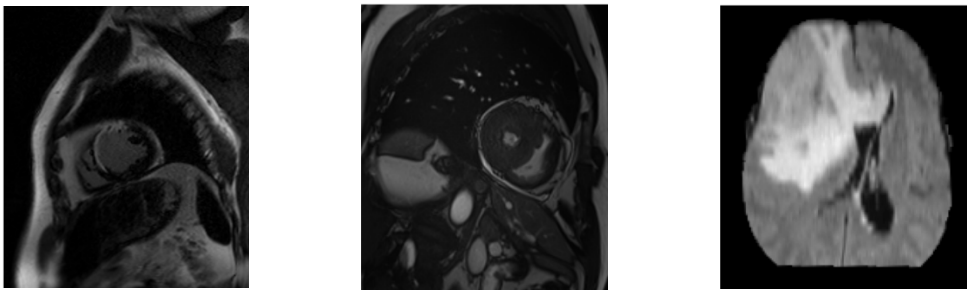
volume becomes a sample, in contrast to 2D methods, which treat volume slices independently.

Although the synthesis method of Chapter 4 showed coherent multi-view results, synthetic results demonstrated a lack of sharpness to be considered realistic by medical experts. This could be improved with new techniques that modern generative models use for high quality synthesis, as for example in [218], where synthesis models are learnt progressively, starting from low resolution and upscaling to high resolution image synthesis.

Our work of Chapter 6 further encourages future extensions to improve the fidelity of reconstructed images by explicitly modelling image texture, which would benefit applications in ultrasound. This can be achieved with the design of more powerful decoders, although how best to maintain the balance between the semantics of the spatial representation, promoted through the thresholding operation, and the quality of the reconstruction is an open question. Texture can be explicitly learned in additional latent variables by using deep feature activations, for example by extracting the Gram matrix as originally proposed for style transfer [128].

Moreover, the applicability of the method in Chapter 6 could be extended in a completely unsupervised setting where no annotated examples are available, or in a zero-shot setting where some annotated examples of other classes are available. Unsupervised segmentation could be possible by careful design of the mask adversarial training, for example with Wasserstein GANs [51] and multi-scale discriminators, and by applying restrictions on the anatomy factor that introduce statistical priors on the shape of the underlying organs. The aim for these constraints would be to achieve a disentanglement within the anatomy factor, with each channel corresponding to a particular organ. In the current methodology, the contents of the anatomy factor are biased by the image's intensities: many channels encode regions of the image with similar pixel values, resembling intensity clustering. Diversity of organ intensities through multimodal learning (as in Chapter 7) in combination with adversarial training and constraints such as minimum description length and connected component analysis would potentially help organ separation in the anatomy factor, and thus unsupervised segmentation.

Disentangled representations, similar to multimodal learning, are also potentially useful to transfer learning, for example for multi-site data of the same modality. In this case, the multi-site variability would be encoded in the modality factor. In the transfer learning scenario, we want to use a trained model on a new dataset with no annotations. The unseen intensity patterns of the new data may potentially affect image disentanglement and therefore segmentation ac-



(a) cardiac infarct **(b)** hypertrophic cardiomyopathy **(c)** brain tumour

Figure 8.1: Different pathology examples that affect appearance (hyperintense infarcted region), size (hypertrophic myocardium), and shape (brain tumour). Images are taken from ERI [2] (Section 2.6.2.5), ACDC [5] (Section 2.6.2.2) and BRATS [16] (Section 2.6.1.2).

curacy. We believe that fine tuning the encoders (or their first few layers) for a small number of epochs using the unsupervised costs would not affect the semantics of the anatomy space and suffice for extending our model to the new data.

We further believe, that there is a big potential in the research of disentangled representations. Currently, only two factors are considered, but explicitly learning hierarchical factors that better capture semantic information (both in terms of anatomical and modality representations), would create structured representations that help generalisation in different applications. For instance, the discovery of pathology factors can have direct application in automatic pathology classification, although different pathologies manifest in different ways. Some pathologies affect the "appearance" of anatomical regions, e.g. cardiac infarct appears as hyperintense regions of the myocardium in LGE, others affect the shape of the organ, e.g. hypertrophic cardiomyopathy results in increased myocardial volume, and others deform the shape of healthy tissues, e.g. due to introduction of cancerous mass. We believe that key in disentangling pathology is maintaining a representation of the healthy anatomy. As such, an anatomy encoder should always extract healthy anatomical features (encouraged by adversarial learning or statistical constraints, e.g. size and shape), with the pathology factor being estimated as the residual representation required to reconstruct the image. Although it might be tempting to encode pathologies that affect intensities in the modality factor, this would contradict the anatomy-modality disentanglement principle. Furthermore, disentangling pathologies that induce deformations, such as brain tumour, is more challenging, since the anatomy encoder should "undo" the deformation, similar to pseudo-healthy synthesis methods [69, 80].

Hierarchical representations that model conditional dependencies between latent variables can also be employed to encourage decoding from coarse to fine-grained details. Indeed, this could also be a step towards solutions requiring less (or weaker) supervision. Learning such an expressive generative model with hierarchical factors requires a different §decoder design, which should generate images by compositionality, i.e. by combining (and colouring) separate objects. Compositional methods in computer vision take advantage of data biases, where the same object appears in different colours and poses. With appropriate adversarial and information theoretic costs, the GAN method in [219] considers a sequential approach, where background, object shape and appearance are generated in a sequence and "stiched" together. The data biases in medical imaging are different: intensity variations between images of the same organ are typically very small, whereas large intensity variations among organs are not guaranteed, and if any, they depend on physical properties, such as the hydrogen concentration. Nevertheless, there is anatomical variability that is correlated to other attributes, e.g. the temporal frames of a cine-MR sequence or the slice position within a 3D volume. We believe that compositional decoders that encode these factors along with traditional anatomy, modality and pathology factors is a promising future direction. Furthermore, a separation between background and foreground anatomy, which in CMR images corresponds to the heart and surrounding organs respectively, is necessary for modelling the dependencies with the temporal and spatial factors. Initially, assuming some supervision on the heart, either strong (segmentation masks) or weak (bounding boxes), a compositional generative model given temporal and spatial coordinates would respectively: synthesise a background anatomy, a healthy heart, a representation of some pathology, and finally would create a composition of the previous given a modality.

Finally, a theoretical characterisation of the disentangling process and precise quantification of the type of information that is captured by each factor is required, in order to fully take advantage and tune disentangled representation methods according to specific learning tasks. This admittedly is more complex in spatial disentanglement than in vectorised latent spaces for which metrics have been recently suggested [220]. An initial investigation has been performed in Chapter 7 by using distance correlation between the anatomy and modality factors, whereas a more thorough analysis examining several biases that affect disentanglement in different computer vision models has been conducted in our article titled "Metrics for Exposing the Biases of Content-Style Disentanglement", which at the time of writing is under review.

References

[1] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, *et al.*, "ISLES 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri," *Medical Image Analysis*, vol. 35, pp. 250–269, 2017.

[2] C. G. Stirrat, S. R. Alam, T. J. MacGillivray, C. D. Gray, M. R. Dweck, J. Raftis, *et al.*, "Ferumoxytol-enhanced magnetic resonance imaging assessing inflammation after myocardial infarction," *Heart*, vol. 103, no. 19, pp. 1528–1535, 2017.

[3] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes, and S. Ourselin, "A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI," *IEEE Transactions on Medical Imaging*, vol. 29, no. 9, pp. 1612–1625, 2010.

[4] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of mri," *Medical Image Analysis*, vol. 31, pp. 77–87, 2016.

[5] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P. Jodoin, "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2514–2525, Nov 2018.

[6] W. Manning and D. Pennell, *Cardiovascular Magnetic Resonance*. Elsevier, 2003.

[7] T. M. Buzug, "Computed tomography from photon statistics to modern cone-beam CT," 2008.

[8] Wikipedia contributors, "Heart – Wikipedia, the free encyclopedia," 2019. [Online; accessed 3-February-2020].

[9] Wikipedia contributors, "Wiggers diagram – Wikipedia, the free encyclopedia," 2019. [Online; accessed 3-February-2020].

[10] S. A. Tsaftaris, X. Zhou, R. Tang, D. Li, and R. Dharmakumar, "Detecting myocardial ischemia at rest with cardiac phase–resolved blood oxygen level–dependent cardiovascular magnetic resonance," *Circulation: Cardiovascular Imaging*, vol. 6, no. 2, pp. 311–319, 2013.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, pp. 1125–1134, 2017.

[12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *ICLR*, 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241, Springer, 2015.

[14] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *AAAI*, pp. 3942–3951, AAAI Press, 2018.

[15] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *CVPR*, pp. 2337–46, 2019.

[16] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[17] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[18] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical Image Analysis*, vol. 35, pp. 18–31, 2017.

[19] S. E. Petersen, P. M. Matthews, F. Bamberg, D. A. Bluemke, J. M. Francis, M. G. Friedrich, P. Leeson, E. Nagel, S. Plein, F. E. Rademakers, *et al.*, "Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of uk biobank-rationale, challenges and approaches," *Journal of Cardiovascular Magnetic Resonance*, vol. 15, no. 1, p. 46, 2013.

[20] X. Zhuang, "Challenges and methodologies of fully automatic whole heart segmentation: a review," *Journal of healthcare engineering*, vol. 4, no. 3, pp. 371–407, 2013.

[21] S. Narayanaswamy, T. B. Paige, J.-W. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr, "Learning disentangled representations with semi-supervised deep generative models," in *NIPS*, pp. 5927–5937, 2017.

[22] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[23] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, pp. 689–696, 2011.

[24] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *CVPR*, pp. 2940–2949, 2016.

[25] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *ECCV*, pp. 172–189, 2018.

[26] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *CVPR*, pp. 10955–10964, 2019.

[27] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *ICML*, pp. 4114–4124, 2019.

[28] T. Joyce and S. Kozerke, "3D medical image synthesis by factorised representation and deformable model learning," in *SASHIMI*, pp. 110–119, Springer, 2019.

[29] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 803–814, 2017.

[30] A. Chartsias, T. Joyce, R. Dharmakumar, and S. A. Tsaftaris, "Adversarial image synthesis for unpaired multi-modal cardiac data," in *SASHIME*, pp. 3–13, Springer International Publishing, 2017.

[31] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Factorised spatial representation learning: Application in semi-supervised myocardial segmentation," in *MICCAI*, (Cham), pp. 490–498, Springer International Publishing, 2018.

[32] A. Chartsias, T. Joyce, G. Papanastasiou, S. Semple, M. Williams, D. E. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangled representation learning in cardiac image analysis," *Medical Image Analysis*, vol. 58, p. 101535, 2019.

[33] A. Chartsias, G. Papanastasiou, C. Wang, C. Stirrat, S. Semple, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Multimodal cardiac segmentation using disentangled representation learning," in *STACOM*, (Cham), pp. 128–137, Springer International Publishing, 2020.

[34] A. Chartsias, G. Papanastasiou, C. Wang, S. Semple, D. Newby, R. Dharmakumar, and S. A. Tsaftaris, "Disentangle, align and fuse for multimodal and zero-shot image segmentation," *arXiv preprint arXiv:1911.04417*, 2019.

[35] P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi, "A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 29, no. 2, pp. 155–195, 2016.

[36] M. Saeed, S. Wagner, M. Wendland, N. Derugin, W. Finkbeiner, and C. Higgins, "Occlusive and reperfused myocardial infarcts: differentiation with mn-dpdp–enhanced mr imaging.," *Radiology*, vol. 172, no. 1, pp. 59–64, 1989.

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, pp. 248–255, 2009.

[38] S. H. Park and K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," *Radiology*, vol. 286, no. 3, pp. 800–809, 2018.

[39] "IXI dataset - brain development." [Online; accessed 3-February-2020].

[40] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer, "CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data," Apr. 2019.

[41] A. E. Kavur, N. S. Gezer, M. Barış, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, S. Aslan, D. Sheet, G. Dovletov, O. Speck, A. Nürnberger, K. H. Maier-Hein, G. B. Akar, G. Ünal, O. Dicle, and M. A. Selver, "CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation," Jan. 2020.

[42] A. E. Kavur, N. S. Gezer, M. Barış, Y. Şahin, S. Özkan, B. Baydar, U. Yüksel, Kılıkçıer, Olut, G. Bozdağı Akar, G. Ünal, O. Dicle, and M. A. Selver, "Comparison of semi-automatic and deep learning based automatic methods for liver segmentation in living liver transplant donors," *Diagnostic and Interventional Radiology*, vol. 26, pp. 11–21, Jan. 2020.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, pp. 2672–80, 2014.

[44] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[45] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *ICML*, vol. 32, pp. 1278–1286, PMLR, 22–24 Jun 2014.

[46] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural networks*, vol. 6, no. 6, pp. 861–867, 1993.

[47] Z. Wang, Q. She, and T. E. Ward, "Generative adversarial networks: A survey and taxonomy," *arXiv preprint arXiv:1906.01529*, 2019.

[48] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *ICLR*, 2016.

[49] A. Jahanian, L. Chai, and P. Isola, "On the "steerability" of generative adversarial networks," *ICLR*, 2020.

[50] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "On the effectiveness of least squares generative adversarial networks," *arXiv:1712.06391*, 2017.

[51] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *ICML*, 2017.

[52] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein GANs," in *NIPS*, pp. 5767–5777, 2017.

[53] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *ICLR*, 2018.

[54] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: Boundary equilibrium generative adversarial networks," *preprint arXiv:1703.10717*, 2017.

[55] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," in *NIPS*, pp. 3308–18, 2017.

[56] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? a large-scale study," in *NIPS*, pp. 700–709, 2018.

[57] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *arXiv preprint arXiv:2001.06937*, 2020.

[58] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[59] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, pp. 2223–2232, 2017.

[60] A. Makhzani, J. Shlens, N. Jaitly, and I. Goodfellow, "Adversarial autoencoders," in *ICLR*, 2016.

[61] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, "Correlational neural networks," *Neural computation*, 2016.

[62] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *CVPR*, pp. 1933–1941, 2016.

[63] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, p. 1153, 2013.

[64] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[65] H. Li, S. Jialin Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *CVPR*, pp. 5400–5409, 2018.

[66] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[67] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "HeMIS: Hetero-modal image segmentation," in *MICCAI*, pp. 469–477, Springer, 2016.

[68] H.-I. Suk, S.-W. Lee, D. Shen, A. D. N. Initiative, *et al.*, "Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis," *NeuroImage*, vol. 101, pp. 569–582, 2014.

[69] C. Bowles, C. Qin, C. Ledig, R. Guerrero, R. Gunn, A. Hammers, E. Sakka, D. A. Dickie, M. V. Hernández, N. Royle, *et al.*, "Pseudo-healthy image synthesis for white matter lesion segmentation," in *SASHIMI*, pp. 87–96, Springer, 2016.

[70] T. Huynh, Y. Gao, J. Kang, L. Wang, P. Zhang, J. Lian, and D. Shen, "Estimating CT image from MRI data using structured random forest and auto-context model," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 174–183, 2016.

[71] A. Jog, A. Carass, S. Roy, D. L. Pham, and J. L. Prince, "Random forest regression for magnetic resonance image synthesis," *Medical Image Analysis*, vol. 35, pp. 475–88, 2017.

[72] S. Roy, Y.-Y. Chou, A. Jog, J. A. Butman, and D. L. Pham, "Patch based synthesis of whole head MR images: Application to EPI distortion correction," in *SASHIMI*, pp. 146–156, Springer, 2016.

[73] D. H. Ye, D. Zikic, B. Glocker, A. Criminisi, and E. Konukoglu, "Modality propagation: coherent synthesis of subject-specific scans with data-driven regularization," in *MICCAI*, pp. 606–613, Springer, 2013.

[74] S. Roy, A. Carass, and J. L. Prince, "Magnetic resonance image example-based contrast synthesis," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2348–63, 2013.

[75] Y. Huang, L. Beltrachini, L. Shao, and A. F. Frangi, "Geometry regularized joint dictionary learning for cross-modality image synthesis in magnetic resonance imaging," in *SASHIMI*, pp. 118–126, Springer, 2016.

[76] R. Vemulapalli, H. Van Nguyen, and S. Kevin Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *ICCV*, pp. 630–638, 2015.

[77] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *MICCAI*, pp. 305–312, Springer, 2014.

[78] H. Van Nguyen, K. Zhou, and R. Vemulapalli, "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *MICCAI*, pp. 677–84, Springer, 2015.

[79] V. Sevetlidis, M. V. Giuffrida, and S. A. Tsaftaris, "Whole image synthesis using a deep encoder-decoder network," in *SASHIMI*, pp. 97–107, Springer, 2016.

[80] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Adversarial pseudo healthy synthesis needs pathology factorization," *MIDL*, 2019.

[81] X. Yang, X. Han, E. Park, S. Aylward, R. Kwitt, and M. Niethammer, "Registration of pathological images," in *SASHIMI*, pp. 97–107, Springer International Publishing, 2016.

[82] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, "Image synthesis in multi-contrast mri with conditional generative adversarial networks," *IEEE Transactions on Medical Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.

[83] K. Armanious, C. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, vol. 79, p. 101684, 2020.

[84] D. Nie, R. Trullo, J. Lian, L. Wang, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with deep convolutional adversarial networks," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 12, pp. 2720–2730, 2018.

[85] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.

[86] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[87] T. Xia, A. Chartsias, S. A. Tsaftaris, A. D. N. Initiative, *et al.*, "Consistent brain ageing synthesis," in *MICCAI*, pp. 750–758, Springer, 2019.

[88] J. M. Wolterink, A. M. Dinkla, M. H. Savenije, P. R. Seevinck, C. A. van den Berg, and I. Išgum, "Deep MR to CT synthesis using unpaired data," in *SASHIMI*, pp. 14–23, Springer, 2017.

[89] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "Cross-modality image synthesis from unpaired data using cyclegan," in *SASHIMI*, pp. 31–41, Springer, 2018.

[90] J. Cai, Z. Zhang, L. Cui, Y. Zheng, and L. Yang, "Towards cross-modal organ translation and segmentation: a cycle-and shape-consistent generative adversarial network," *Medical Image Analysis*, vol. 52, pp. 174–184, 2019.

[91] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, p. 101552, 2019.

[92] O. Oktay *et al.*, "Multi-input cardiac image super-resolution using convolutional neural networks," in *MICCAI*, pp. 246–254, Springer, 2016.

[93] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. A. Cook, A. De Marvao, T. Dawes, D. P. O'Regan, *et al.*, "Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 384–395, 2017.

[94] Y. Huang, L. Shao, and A. F. Frangi, "Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding," in *CVPR*, pp. 6070–6079, 2017.

[95] J. Ossenberg-Engels and V. Grau, "Conditional generative adversarial networks for the prediction of cardiac contraction from individual frames," in *STACOM*, pp. 109–118, Springer, 2019.

[96] X. Tao, H. Wei, W. Xue, and D. Ni, "Segmentation of multimodal myocardial images using shape-transfer GAN," in *STACOM*, pp. 271–279, Springer, 2020.

[97] H. Qiu, W. Bai, and D. Rueckert, "Unsupervised multi-modal style transfer for cardiac MR segmentation," in *STACOM*, p. 209, Springer.

[98] G. Valvano, A. Chartsias, A. Leo, and S. A. Tsaftaris, "Temporal consistency objectives regularize the learning of disentangled representations," in *DART*, pp. 11–19, Springer, 2019.

[99] N. Cordier, H. Delingette, M. Lê, and N. Ayache, "Extended modality propagation: Image synthesis of pathological cases," *IEEE Transactions on Medical Imaging*, vol. 35, no. 12, pp. 2598–2608, 2016.

[100] Y. Wang, L. Zhou, B. Yu, L. Wang, C. Zu, D. S. Lalush, W. Lin, X. Wu, J. Zhou, and D. Shen, "3D auto-context-based locality adaptive multi-modality GANs for PET synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 6, pp. 1328–1339, 2018.

[101] A. Sharma and G. Hamarneh, "Missing mri pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Transactions on Medical Imaging*, 2019.

[102] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye, "CollaGAN: Collaborative gan for missing image data imputation," in *CVPR*, pp. 2487–2496, 2019.

[103] H. Li, J. C. Paetzold, A. Sekuboyina, F. Kofler, J. Zhang, J. S. Kirschke, B. Wiestler, and B. Menze, "DiamondGAN: Unified multi-modal generative adversarial networks for mri sequences synthesis," in *MICCAI*, pp. 795–803, Springer, 2019.

[104] M. Yurt, S. U. H. Dar, A. Erdem, E. Erdem, and T. Çukur, "mustGAN: Multi-stream generative adversarial networks for mr image synthesis," *arXiv preprint arXiv:1909.11504*, 2019.

[105] G. Chen and S. N. Srihari, "Generalized K-fan multimodal deep model with shared representations," *arXiv preprint arXiv:1503.07906*, 2015.

[106] V. Vukotić, C. Raymond, and G. Gravier, "Bidirectional Joint Representation Learning with Symmetrical Deep Neural Networks for Multimodal and Crossmodal Applications," in *ICMR*, (New York, New York, USA), pp. 343–346, ACM Press, 2016.

[107] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," in *cvpr*, pp. 3020–3028, 2017.

[108] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[109] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in *CVPR*, pp. 5447–5455, 2017.

[110] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, pp. 1–24, 2019.

[111] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *NIPS*, pp. 343–351, 2016.

[112] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *CVPR*, pp. 7167–7176, 2017.

[113] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[114] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *ICLR*, 2017.

[115] A. Royer, K. Bousmalis, S. Gouws, F. Bertsch, I. Mosseri, F. Cole, and K. Murphy, "Xgan: Unsupervised image-to-image translation for many-to-many mappings," in *Domain Adaptation for Visual Understanding*, pp. 33–49, Springer, 2020.

[116] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *NIPS*, pp. 700–708, 2017.

[117] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," *ICLR workshop*, 2015.

[118] H. Kim and A. Mnih, "Disentangling by factorising," *ICML*, 2018.

[119] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *NIPS*, pp. 2172–2180, 2016.

[120] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *NIPS*, pp. 613–621, 2016.

[121] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," *CVPR*, 2018.

[122] C. Donahue, Z. C. Lipton, A. Balsubramani, and J. McAuley, "Semantically decomposing the latent spaces of generative adversarial networks," in *ICLR*, 2018.

[123] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun, "Disentangling factors of variation in deep representation using adversarial training," in *NIPS*, pp. 5040–5048, 2016.

[124] A. Szabó, Q. Hu, T. Portenier, M. Zwicker, and P. Favaro, "Challenges in disentangling independent factors of variation," in *ICLR Workshop*, 2018.

[125] Q. Hu, A. Szabó, T. Portenier, M. Zwicker, and P. Favaro, "Disentangling factors of variation by mixing them," *CVPR*, 2018.

[126] C. Biffi, O. Oktay, G. Tarroni, W. Bai, A. De Marvao, G. Doumou, M. Rajchl, R. Bedair, S. Prasad, S. Cook, D. O'Regan, and D. Rueckert, "Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling," in *MICCAI*, (Cham), pp. 464–471, Springer International Publishing, 2018.

[127] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren, "Scalable multimodal convolutional networks for brain tumour segmentation," in *MICCAI*, (Cham), pp. 285–293, Springer International Publishing, 2017.

[128] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, pp. 2414–2423, 2016.

[129] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *NIPS*, pp. 465–476, 2017.

[130] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented CycleGAN: Learning many-to-many mappings from unpaired data," *ICML*, 2018.

[131] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *ECCV*, vol. 11205, pp. 36–52, Springer International Publishing, 2018.

[132] P. Esser, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," in *CVPR*, pp. 8857–8866, 2018.

[133] B. Lu, J.-C. Chen, and R. Chellappa, "UID-GAN: Unsupervised image deblurring via disentangled representations," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2019.

[134] L. Ma, X. Jia, S. Georgoulis, T. Tuytelaars, and L. V. Gool, "Exemplar guided unsupervised image-to-image translation," in *ICLR*, 2019.

[135] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, "TransGaGa: Geometry-aware unsupervised image-to-image translation," *CVPR*, 2019.

[136] W. Wu, K. Cao, C. Li, C. Qian, and C. C. Loy, "Disentangling content and style via unsupervised geometry distillation," *ICLR workshop*, 2019.

[137] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *IPMI*, pp. 249–261, Springer, 2019.

[138] Q. Meng, N. Pawlowski, D. Rueckert, and B. Kainz, "Representation disentanglement for multi-task learning with application to fetal ultrasound," in *PIPPI*, pp. 47–55, Springer, 2019.

[139] Q. Meng, D. Rueckert, and B. Kainz, "Learning cross-domain generalizable features by representation disentanglement," *arXiv preprint arXiv:2003.00321*, 2020.

[140] A. Ben-Cohen, R. Mechrez, N. Yedidia, and H. Greenspan, "Improving cnn training using disentanglement for liver lesion classification in ct," in *EMBC*, pp. 886–889, IEEE, 2019.

[141] H. Liao, W.-A. Lin, S. K. Zhou, and J. Luo, "Adn: Artifact disentanglement network for unsupervised metal artifact reduction," *IEEE Transactions on Medical Imaging*, 2019.

[142] S. Liu, E. Gibson, S. Grbic, Z. Xu, A. A. A. Setio, J. Yang, B. Georgescu, and D. Comaniciu, "Decompose to manipulate: Manipulable object synthesis in 3D medical images with structured image decomposition," *arXiv:1812.01737*, 2018.

[143] K. Li, L. Yu, S. Wang, and P.-A. Heng, "Unsupervised retina image synthesis via disentangled representation learning," in *SASHIMI*, pp. 32–41, Springer, 2019.

[144] J. Yang, N. C. Dvornek, F. Zhang, J. Zhuang, J. Chapiro, M. Lin, and J. S. Duncan, "Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation," *ICCV*, 2019.

[145] J. Yang, N. C. Dvornek, F. Zhang, J. Chapiro, M. Lin, and J. S. Duncan, "Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation," in *MICCAI*, pp. 255–263, Springer, 2019.

[146] C. Chen, Q. Dou, Y. Jin, H. Chen, J. Qin, and P.-A. Heng, "Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion," in *MICCAI*, pp. 447–456, Springer, 2019.

[147] Y. Lyu, H. Liao, H. Zhu, and S. K. Zhou, "Joint unsupervised learning for the vertebra segmentation, artifact reduction and modality translation of cbct images," *arXiv preprint arXiv:2001.00339*, 2020.

[148] World Health Organization, "Cardiovascular Diseases," 2017. [Online; accessed 3-February-2020].

[149] M. A. Fogel, "Use of ejection fraction (or lack thereof), morbidity/mortality and heart failure drug trials: a review," *International journal of cardiology*, vol. 84, no. 2-3, pp. 119–132, 2002.

[150] C. Rickers, N. M. Wilke, M. Jerosch-Herold, S. A. Casey, P. Panse, N. Panse, J. Weil, A. G. Zenovich, and B. J. Maron, "Utility of cardiac magnetic resonance imaging in the diagnosis of hypertrophic cardiomyopathy," *Circulation*, vol. 112, no. 6, pp. 855–861, 2005.

[151] T. A. Ngo, Z. Lu, and G. Carneiro, "Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance," *Medical Image Analysis*, vol. 35, pp. 159–171, 2017.

[152] M. R. Avendi, A. Kheradvar, and H. Jafarkhani, "Automatic segmentation of the right ventricle from cardiac mri using a learning-based approach," *Magnetic Resonance in Medicine*, 2017.

[153] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis mri," *preprint arXiv:1604.00494*, 2016.

[154] J. Lieman-Sifry, M. Le, F. Lau, S. Sall, and D. Golden, "FastVentricle: cardiac segmentation with enet," in *FIMH*, pp. 127–138, Springer, 2017.

[155] L. K. Tan, Y. M. Liew, E. Lim, and R. A. McLaughlin, "Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine mr sequences," *Medical Image Analysis*, vol. 39, pp. 78–86, 2017.

[156] R. P. Poudel, P. Lamata, and G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *RAMBO, HVSMR*, pp. 83–94, Springer, 2016.

[157] D. M. Vigneault, W. Xie, C. Y. Ho, D. A. Bluemke, and J. A. Noble, "Ω-net (omega-net): Fully automatic, multi-view cardiac mr detection, orientation, and segmentation with deep neural networks," *Medical Image Analysis*, vol. 48, pp. 95 – 106, 2018.

[158] W. Bai, M. Sinclair, G. Tarroni, O. Oktay, M. Rajchl, G. Vaillant, A. M. Lee, N. Aung, E. Lukaschuk, M. M. Sanghvi, *et al.*, "Automated cardiovascular magnetic resonance image analysis with fully convolutional networks," *Journal of Cardiovascular Magnetic Resonance*, vol. 20, no. 1, p. 65, 2018.

[159] C. Zotti, Z. Luo, A. Lalande, and P.-M. Jodoin, "Convolutional neural network with shape prior applied to cardiac mri segmentation," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 1119–1128, 2018.

[160] C. Qin, W. Bai, J. Schlemper, S. E. Petersen, S. K. Piechnik, S. Neubauer, and D. Rueckert, "Joint motion estimation and segmentation from undersampled cardiac mr image," in *Machine Learning for Medical Image Reconstruction* (F. Knoll, A. Maier, and D. Rueckert, eds.), (Cham), pp. 55–63, Springer International Publishing, 2018.

[161] W. Bai, H. Suzuki, C. Qin, G. Tarroni, O. Oktay, P. M. Matthews, and D. Rueckert, "Recurrent neural networks for aortic image sequence segmentation with sparse annotations," in *MICCAI*, (Cham), pp. 586–594, Springer International Publishing, 2018.

[162] A. Mortazi, R. Karim, K. Rhode, J. Burt, and U. Bagci, "CardiacNET: segmentation of left atrium and proximal pulmonary veins from mri using multi-view cnn," in *MICCAI*, pp. 377–385, Springer, 2017.

[163] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P.-A. Heng, "3D deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, vol. 41, pp. 40–54, 2017.

[164] Q. Zheng, H. Delingette, N. Duchateau, and N. Ayache, "3-D consistent and robust segmentation of cardiac images by deep learning with spatial propagation," *IEEE Transactions on Medical Imaging*, vol. 37, pp. 2137–2148, Sept 2018.

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

[166] V. M. Campello, C. Martín-Isla, C. Izquierdo, S. E. Petersen, M. A. G. Ballester, and K. Lekadir, "Combining multi-sequence and synthetic images for improved segmentation of late gadolinium enhancement cardiac mri," in *STACOM*, (Cham), pp. 290–299, Springer International Publishing, 2020.

[167] Y. Huo, Z. Xu, H. Moon, S. Bao, A. Assad, T. K. Moyo, M. R. Savona, R. G. Abramson, and B. A. Landman, "SynSeg-Net: Synthetic segmentation without target modality ground truth," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 1016–25, 2019.

[168] B. Ly, H. Cochet, and M. Sermesant, "Style data augmentation for robust segmentation of multi-modality cardiac MRI," in *STACOM*, pp. 197–208, Springer, 2020.

[169] Y. Liu, W. Wang, K. Wang, C. Ye, and G. Luo, "An automatic cardiac segmentation framework based on multi-sequence MR image," in *STACOM*, pp. 220–7, Springer, 2020.

[170] C. Chen, Q. Dou, H. Chen, J. Qin, and P.-A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," *AAAI*, 2019.

[171] W. Bai, O. Oktay, M. Sinclair, H. Suzuki, M. Rajchl, G. Tarroni, B. Glocker, A. King, P. M. Matthews, and D. Rueckert, "Semi-supervised learning for network-based cardiac MR image segmentation," in *MICCAI*, (Cham), pp. 253–260, Springer, 2017.

[172] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *MICCAI*, pp. 408–416, Springer, 2017.

[173] C. Baur, S. Albarqouni, and N. Navab, "Semi-supervised deep learning for fully convolutional networks," in *MICCAI*, pp. 311–319, Springer, 2017.

[174] Z. Zhao, L. Yang, H. Zheng, I. H. Guldner, S. Zhang, and D. Z. Chen, "Deep learning based instance segmentation in 3D biomedical images using weak annotation," in *MICCAI*, (Cham), pp. 352–360, Springer International Publishing, 2018.

[175] D. Nie, Y. Gao, L. Wang, and D. Shen, "ASDNet: Attention based semi-supervised deep networks for medical image segmentation," in *MICCAI*, pp. 370–8, Springer, 2018.

[176] H. Kervadec, J. Dolz, E. Granger, and I. B. Ayed, "Curriculum semi-supervised segmentation," in *MICCAI*, pp. 568–576, Springer, 2019.

[177] W. Bai, C. Chen, G. Tarroni, J. Duan, F. Guitton, S. E. Petersen, Y. Guo, P. M. Matthews, and D. Rueckert, "Self-supervised learning for cardiac mr image segmentation by anatomical position prediction," in *MICCAI*, pp. 541–549, Springer, 2019.

[178] P. Luc, C. Couprie, S. Chintala, and J. Verbeek, "Semantic segmentation using adversarial networks," *NIPS Workshop*, 2016.

[179] N. Souly, C. Spampinato, and M. Shah, "Semi and weakly supervised semantic segmentation using generative adversarial network," *arXiv:1703.09695*, 2017.

[180] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," in *IPMI*, pp. 29–41, Springer, 2019.

[181] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[182] G. van Tulder and M. de Bruijne, "Learning cross-modality representations from multi-modal images," *IEEE Transactions Medical Imaging*, vol. 38, no. 2, pp. 638–648, 2019.

[183] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. B. Ayed, "HyperDenseNet: A hyper-densely connected CNN for multi-modal image segmentation," *IEEE Transactions Medical Imaging*, vol. 38, no. 5, pp. 1116–1126, 2019.

[184] F. Mahmood, Z. Yang, T. Ashley, and N. J. Durr, "Multimodal densenet," *arXiv preprint arXiv:1811.07407*, 2018.

[185] R. Couronne, M. Louis, and S. Durrleman, "Longitudinal autoencoder for multi-modal disease progression modelling," 2019.

[186] K.-L. Tseng, Y.-L. Lin, W. Hsu, and C.-Y. Huang, "Joint sequence learning and cross-modality convolution for 3D biomedical segmentation," in *CVPR*, pp. 6393–6400, 2017.

[187] H. Chen, Y. Qi, Y. Yin, T. Li, X. Liu, X. Li, G. Gong, and L. Wang, "Mmfnet: A multi-modality mri fusion network for segmentation of nasopharyngeal carcinoma," *Neurocomputing*, 2020.

[188] M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, "Spatial transformer networks," in *NIPS*, pp. 2017–2025, 2015.

[189] V. Valindria, N. Pawlowski, M. Rajchl, I. Lavdas, E. Aboagye, A. Rockall, D. Rueckert, and B. Glocker, "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI," in *WACV*, pp. 547–556, IEEE, 2018.

[190] J. Liu, H. Xie, S. Zhang, and L. Gu, "Multi-sequence myocardium segmentation with cross-constrained shape and neural network-based initialization," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 49–57, 2019.

[191] D. Wei, Y. Sun, P. Chai, A. Low, and S. H. Ong, "Myocardial segmentation of late gadolinium enhanced MR images by propagation of contours from cine MR images," in *MICCAI*, pp. 428–435, Springer, 2011.

[192] X. Zhuang, "Multivariate mixture model for myocardial segmentation combining multi-source images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 12, pp. 2933–2946, 2019.

[193] X. Wang, S. Yang, M. Tang, Y. Wei, X. Han, L. He, and J. Zhang, "SK-Unet: an improved u-net model with selective kernel for the segmentation of multi-sequence cardiac mr," in *STACOM*, pp. 246–253, Springer, 2020.

[194] S. Vesal, N. Ravikumar, and A. Maier, "Automated multi-sequence cardiac mri segmentation using supervised domain adaptation," *STACOM*, 2019.

[195] J. Chen, H. Li, J. Zhang, and B. Menze, "Adversarial convolutional networks with weak domain-transfer for multi-sequence cardiac MR images segmentation," *arXiv preprint arXiv:1908.09298*, 2019.

[196] H. Roth, W. Zhu, D. Yang, Z. Xu, and D. Xu, "Cardiac segmentation of LGE MRI with noisy labels," in *STACOM*, pp. 228–236, Springer, 2020.

[197] R. Zheng, X. Zhao, X. Zhao, and H. Wang, "Deep learning based multi-modal cardiac MR image segmentation," in *STACOM*, (Cham), pp. 263–270, Springer, 2020.

[198] A. F. Frangi, S. A. Tsaftaris, and J. L. Prince, "Simulation and synthesis in medical imaging," *IEEE Transactions on Medical Imaging*, vol. 37, no. 3, pp. 673–679, 2018.

[199] G. van Tulder and M. de Bruijne, "Why does synthesized data improve multi-sequence classification?," in *MICCAI*, pp. 531–538, Springer, 2015.

[200] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *LABELS*, pp. 179–87, Springer, 2016.

[201] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR*, 2014.

[202] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[203] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *NIPS*, pp. 3320–3328, 2014.

[204] C. Chu, A. Zhmoginov, and M. Sandler, "CycleGAN: a master of steganography," *NIPS workshop on Machine Deception*, 2017.

[205] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[206] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *3DV*, pp. 565–571, 2016.

[207] 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, p. 65, Sep 2018.

[208] M. Bevilacqua, R. Dharmakumar, and S. A. Tsaftaris, "Dictionary-driven ischemia detection from cardiac phase-resolved myocardial BOLD MRI at rest," *IEEE Transactions on Medical Imaging*, vol. 35, pp. 282–293, Jan 2016.

[209] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *NIPS Workshop on Learning Disentangled Representations*, 2018.

[210] H. W. Kim, A. Farzaneh-Far, and R. J. Kim, "Cardiovascular magnetic resonance in patients with myocardial infarction: current and emerging applications," *Journal of the American College of Cardiology*, vol. 55, no. 1, pp. 1–16, 2009.

[211] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," *ICLR*, 2017.

[212] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, pp. 1321–1330, JMLR.org, 2017.

[213] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *UAI*, 2018.

[214] N. J. Tustison, Y. Yang, and M. Salerno, "Advanced normalization tools for cardiac motion correction," in *STACOM*, pp. 3–12, Springer, 2015.

[215] I. Oksuz, J. Clough, B. Ruijsink, E. Puyol-Antón, A. Bustin, G. Cruz, C. Prieto, D. Rueckert, A. P. King, and J. A. Schnabel, "Detection and correction of cardiac MRI motion artefacts during reconstruction from k-space," in *MICCAI*, pp. 695–703, Springer, 2019.

[216] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[217] P. Vepakomma, O. Gupta, A. Dubey, and R. Raskar, "Reducing leakage in distributed deep learning for sensitive health data," *ICLR*, 2019.

[218] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *ICLR*, 2018.

[219] K. K. Singh, U. Ojha, and Y. J. Lee, "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *CVPR*, pp. 6490–6499, 2019.

[220] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," *arXiv preprint arXiv:1908.09961*, 2019.