THE UNIVERSITY *of* EDINBURGH

Representation Learning for Generalisation in Medical Image Analysis

Xiao Liu

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
March 2023

Abstract

To help diagnose, treat, manage, prevent and predict diseases, medical image analysis plays an increasingly crucial role in modern health care. In particular, using machine learning (ML) and deep learning (DL) techniques to process medical imaging data such as MRI, CT and X-Rays scans has been a research hot topic. Accurate and generalisable medical image segmentation using ML and DL is one of the most challenging medical image analysis tasks. The challenges are mainly caused by two key reasons: **a)** the variations of data statistics across different clinical centres or hospitals, and **b)** the lack of extensive annotations of medical data.

To tackle the above challenges, one of the best ways is to learn disentangled representations. Learning disentangled representations aims to separate out, or disentangle, the underlying explanatory generative factors into disjoint subsets. Importantly, disentangled representations can be efficiently learnt from raw training data with limited annotations. Although, it is evident that learning disentangled representations is well suited for the challenges, there are several open problems in this area. First, there is no work to systematically study how much disentanglement is achieved with different learning and design biases and how different biases affect the task performance for medical data. Second, the benefit of leveraging disentanglement to design models that generalise well on new data has not been well studied especially in medical domain. Finally, the independence prior for disentanglement is a too strong assumption that does not approximate well the true generative factors. According to these problems, this thesis focuses on understanding the role of disentanglement in medical image analysis, measuring how different biases affect disentanglement and the task performance, and then finally using disentangled representations to improve generalisation performance and exploring better representations beyond disentanglement.

In the medical domain, content-style disentanglement is one of the most effective frameworks to learn disentangled presentations. It disentangles and encodes image "content" into a spatial tensor, and image appearance or "style" into a vector that contains information on imaging characteristics. Based on an extensive review of disentanglement, I conclude that it is unclear how different design and learning biases affect the performance of content-style disentanglement methods. Hence, two metrics are proposed to measure the degree of content-style disentanglement by evaluating the informativeness and correlation of representations. By modifying the

design and learning biases in three popular content-style disentanglement models, the degree of disentanglement and task performance of different model variants have been evaluated. A key conclusion is that there exists a sweet spot between task performance and the degree of disentanglement; achieving this sweet spot is the key to design disentanglement models.

Generalising deep models to new data from new centres (termed here domains) remains a challenge. This is largely attributed to shifts in data statistics (domain shifts) between source and unseen domains. With the findings of aforementioned disentanglement metrics study, I design two content-style disentanglement approaches for generalisation. First, I propose two data augmentation methods that improve generalisation. The Resolution Augmentation method generates more diverse data by rescaling images to different resolutions. Subsequently, the Factor-based Augmentation method generates more diverse data by projecting the original samples onto disentangled latent spaces, and combining the learned content and style factors from different domains. To learn more generalisable representations, I integrate gradient-based meta-learning in disentanglement. Gradient-based meta-learning splits the training data into meta-train and meta-test sets to simulate and handle the domain shifts during training, which has shown superior generalisation performance. Considering limited annotations of data, I propose a novel semi-supervised meta-learning framework with disentanglement. I explicitly model the representations related to domain shifts. Disentangling the representations and combining them to reconstruct the input image, allows unlabeled data to be used to better approximate the true domain shifts within a meta-learning setting.

Humans can quickly learn to accurately recognise anatomy of interest from medical images with limited guidance. Such recognition ability can easily generalise to new images from different clinical centres and new tasks in other contexts. This rapid and generalisable learning ability is mostly due to the compositional structure of image patterns in the human brain, which is less incorporated in the medical domain. In this thesis, I explore how compositionality can be applied to learning more interpretable and generalisable representations. Overall, I propose that the ground-truth generative factors that generate the medical images satisfy the compositional equivariance property. Hence, a good representation that approximates well the ground-truth factor has to be compositionally equivariant. By modelling the compositional representations with the learnable von-Mises-Fisher kernels, I explore how different design and learning biases can be used to enforce the representations to be more compositionally equivariant under different learning settings.

Overall, this thesis creates new avenues for further research in the area of generalisable representation learning in medical image analysis, which we believe are key to more generalised machine learning and deep learning solutions in healthcare. In particular, the proposed metrics can be used to guide future work on designing better content-style frameworks. The disentanglement-based meta-learning approach sheds light on leveraging meta-learning for better model generalisation in a low-data regime. Finally, compositional representation learning we believe will play an increasingly important role in designing more generalisable and interpretable models in the future.

Lay Summary

Medical image analysis is crucial in modern healthcare for diagnosing, treating, managing, preventing and predicting diseases. Artificial intelligence (AI) techniques can be used for accurate and automatic medical image analysis. However, using AI is challenging for real-world medical applications due to the differences of data collected from different hospitals and the lack of labels of the medical imaging data. AI solutions that can be used across different hospitals are called generalisable AI. This thesis focuses on developing generalisable AI, taking advantage of the large amount of unlabeled medical imaging data. The key technology in this thesis is extracting or learning representative information (termed representations) that is useful and generalisable for the medical image analysis tasks. In this thesis, two metrics are first introduced to evaluate how good the representations of different AI models are. Then, different approaches are proposed to improve the generalisation ability of the representations of AI models. Eventually, inspired by the recognition process in the human brain, a new framework is proposed to learn more generalisable representations. Overall, this thesis creates new avenues for further research in the area of generalisable representation learning in medical image analysis, which we believe are key to more generalised AI solutions in healthcare.

Declaration of Originality

I hereby declare that the research recorded in this thesis and the thesis itself were composed and originated entirely by myself, unless explicitly stated and acknowledged, in the School of Engineering at the University of Edinburgh.

Xiao Liu

Acknowledgements

PhD is a long journey, but it is also a short journey in life. In the past few years, I have met many people who have helped me a lot. Some left Edinburgh after a short time, while others stayed with us for a longer period. I would like to take this opportunity to thank everyone who has played a role in my journey.

I must first thank my principal supervisor, Prof. Sotirios A. Tsaftaris. I graduated from the University of Edinburgh as an undergraduate in 2017 and had a gap year in Xiamen University in 2018. When I was about to return to Edinburgh to continue my studies in the MSc Signal Processing and Communications program, I wrote an email to Sotos about working with him for my MSc project. I was really impressed that on Sotos' personal website, he wrote, "*While I have contributed to several research domains, my mission is via image analysis to help diagnose and understand diseases and provide food for everyone.*". He has been working hard towards this mission for many years, which also inspired and motivated me a lot. I was fortunate to continue my PhD with Sotos after completing my MSc. Three years ago, I was very naive, immature, and unprofessional. I thank Sotos for all the hours he spent teaching me not only how to do research but also how to pursue the goals in my professional career. COVID has affected us a lot in the last few years, and I really appreciate all the support that Sotos provided me with during these challenging times.

I would also like to thank my assistant supervisor, Dr. Alison Q. O'Neil. I am grateful for all the efforts Alison spent on helping me with my research. I will never forget the chats we had about work, life, and my future career. I particularly appreciate that Alison offered me the job opportunity in her team, which helped minimise the uncertainty and anxiety after my PhD.

I want to thank Dr. Javier Escudero Rodriguez, my personal tutor in my MSc and one of the assistant supervisors in my PhD, for providing me with many helpful suggestions on my research and my future career. I also want to thank Prof. Michael Davies and Dr. Wenjia Bai for being the examiners for my thesis.

I would like to thank my colleagues, collaborators, and friends for all the memorable moments during my PhD. Thank you, Agis, Andrei, Valerio, Tian, Greg, Gabriele, Spiros, Marija, Pedro, Victor, Jana, Nikolaos, Haochuan, John, Yuyang, Feng, Fasih, Kostas, Connor, Christopher,

and Ruolin.

I also thank all my friends outside of the PhD. It is my fortune to have you all as friends in my life. Thank you, Xin Liu, Wanhong Liu, Haochen Li, Shuning Xu, Lina Qiu, Haokun Wang, Yulin Geng, Yinhuan Dong, Zhixi Zhang.

I would particularly like to thank my beloved wife, Qisi Zhang. I am very fortunate to have met you in my life. You saved me when I was going through the worst time in my life. You helped me to be strong, always backing me up. Without your support, I would not have been able to achieve this. I love you, my dear wife, and will spend the rest of my life loving you.

Finally, I thank my family for all the support and love. Thank you – my brother, my sisters, my parents. At the current moment, my mom is still in the hospital, suffering from long COVID and many other medical issues. This reminds me that I was born in an underdeveloped city where medical resources were very limited. I remember that the only doctor in the town only graduated from high school and had very limited medical knowledge. My mom's bad health state is mainly caused by such limited and unfair medical support, which motivates me to use artificial intelligence technology to help everyone get high-quality and affordable healthcare in the world, especially in underdeveloped countries. Thank you, mom and dad. Without your hard work in the past decades, I would never be able to finish this thesis.

Xiao Liu

26/3/2023

List of Publications

Thesis publications:

- **Liu, X.**, Sanchez, P., Thermos, S., O'Neil, A. and Tsaftaris, S.A., 2023. Compositionally Equivariant Representation Learning. IEEE Transactions on Medical Imaging (under review).

  Author contributions.

  Liu, X.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Sanchez, P.: conceptualisation, discussion, analysis, validation, writing - review & editing;

  Thermos, S.: conceptualisation, discussion, analysis, validation, writing - review & editing;

  O'Neil, A.Q.: conceptualisation, resources, supervision, validation, writing - review & editing.

  Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.

  Article presented in Chapter 6.

- **Liu, X.**, Thermos, S., Sanchez, P., O'Neil, A. and Tsaftaris, S.A., 2022. vMFNet: Compositionality Meets Domain-generalised Segmentation. In International Conference on Medical Image Computing and Computer Assisted Intervention 2022.

  Author contributions.

  Liu, X.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Thermos, S.: conceptualisation, discussion, analysis, validation, writing - review & editing;

  Sanchez, P.: conceptualisation, discussion, analysis, validation, writing - review & editing;

O'Neil, A.Q.: conceptualisation, resources, supervision, validation, writing - review & editing.

Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.

Article presented in Chapter 6.

- **Liu, X.**, Thermos, S., O'Neil, A. and Tsaftaris, S.A., 2021. Semi-supervised Meta-learning with Disentanglement for Domain-generalised Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2021.

  Author contributions.

  Liu, X.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Thermos, S.: conceptualisation, discussion, analysis, investigation, validation, writing - review & editing;

  O'Neil, A.Q.: conceptualisation, resources, supervision, validation, writing - review & editing.

  Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.

  Article presented in Chapter 5.

- **Liu, X.**, Thermos, S., Chartsias, A., O'Neil, A. and Tsaftaris, S.A., 2020. Disentangled Representations for Domain-generalized Cardiac Segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 187-195). Springer, Cham.

  Author contributions.

  Liu, X.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Thermos, S.: conceptualisation, investigation, validation, writing - original draft, writing - review & editing;

  Chartsias, A.: conceptualisation, investigation, validation, writing - original draft, writing - review & editing;

  O'Neil, A.Q.: conceptualisation, resources, supervision, validation, writing - review &

editing.

Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.

Article presented in Chapter 5.

- **Liu, X.***, Thermos, S.*, Valvano, G.*, Chartsias, A., O'Neil, A. and Tsaftaris, S.A., 2021. Measuring the Biases and Effectiveness of Content-Style Disentanglement. British Machine Vision Conference 2021. *Equal contribution.

  Author contributions.

  Liu, X.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Thermos, S.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Valvano, G.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  Chartsias, A.: conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualisation, writing - original draft, writing - review & editing;

  O'Neil, A.Q.: conceptualisation, resources, supervision, validation, writing - review & editing.

  Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, validation, writing - review & editing.

  Article presented in Chapter 4.

- **Liu, X.***, Sanchez, P.*, Thermos, S.*, O'Neil, A.Q. and Tsaftaris, S.A., 2022. Learning Disentangled Representations in the Imaging Domain. Medical Image Analysis, p.102516. *Equal contribution.

  Author contributions.

  Liu, X.: conceptualisation, investigation, methodology, project administration, writing - original draft, writing - review & editing;

  Sanchez, P.: conceptualisation, investigation, methodology, project administration, writing - original draft, writing - review & editing;

Thermos, S.: conceptualisation, investigation, methodology, project administration, writing - original draft, writing - review & editing;
O'Neil, A.Q.: conceptualisation, resources, supervision, writing - review & editing.
Tsaftaris, S.A.: conceptualisation, funding acquisition, resources, supervision, writing - review & editing.
Article presented in Chapter 3.

Other publications:

- Sanchez, P., **Liu, X.**, O'Neil, A.Q. and Tsaftaris, S.A., 2023. Diffusion Models for Causal Discovery via Topological Ordering. In International Conference on Learning Representations.

- **Liu, X.**, Thermos, S., Sanchez, P., O'Neil, A.Q. and Tsaftaris, S.A., 2023, February. HSIC-InfoGAN: Learning Unsupervised Disentangled Representations by Maximising Approximated Mutual Information. In Medical Applications with Disentanglements: First MICCAI Workshop, MAD 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings (pp. 15-21). Cham: Springer Nature Switzerland.

- Fragemann, J., Ardizzone, L., **Liu, X.**, Tsaftaris, S.A., Egger, J. and Kleesiek, J., 2023. Review of Disentanglement Approaches for Medical Applications: Towards Solving the *Gordian Knot* of Generative Models in Healthcare. ACM Computing Surveys (under review).

- Fragemann, J., **Liu, X.**, Li, J., Tsaftaris, S.A., Egger, J. and Kleesiek, J., 2023. Applying Disentanglement in the Medical Domain: An Introduction for the MAD Workshop. In Medical Applications with Disentanglements: First MICCAI Workshop, MAD 2022, Held in Conjunction with MICCAI 2022, Singapore, September 22, 2022, Proceedings (pp. 3-11). Cham: Springer Nature Switzerland.

- Campello, V.M., Xia, T., **Liu, X.**, Sanchez, P., Martín-Isla, C., Petersen, S.E., Seguí, S., Tsaftaris, S. and Lekadir, K., 2022. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. Frontiers in cardiovascular medicine, p.2693.

- Su, R.*, **Liu, X.*** and Tsaftaris, S.A., 2022. Why patient data cannot be easily forgotten?. In International Conference on Medical Image Computing and Computer Assisted Intervention 2022. *Equal contribution.

- Sanchez, P., Kascenas, A., **Liu, X.**, O'Neil, A.Q. and Tsaftaris, S.A., 2022. What is Healthy? Generative Counterfactual Diffusion for Lesion Localization. In MICCAI Workshop on Deep Generative Models (pp. 34-44). Springer, Cham.

- Thermos, S., **Liu, X.**, O'Neil, A. and Tsaftaris, S.A., 2021. Controllable cardiac synthesis via disentangled anatomy arithmetic. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2021.

- Campello, V.M., ..., **Liu X.**, Tsaftaris, S.A., ..., Lekadir K. 2021. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. IEEE Transactions on Medical Imaging.

- **Liu, X.** and Tsaftaris, S.A., 2020. Have you forgotten? A method to assess if machine learning models have forgotten data. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 95-105). Springer, Cham.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

AE	Auto-Encoder
AI	Artificial Intelligence
ACM	Association for Computing Machinery
ACDC	Automated Cardiac Diagnosis Challenge
AdaIN	Adaptive Instance Normalisation
BN	Batch Normalisation
BYOL	Bootstrap Your Own Latent
BiGAN	Bidirectional Generative Adversarial Network
C	Content
CT	Computed Tomography
CMR	Cardiac Magnetic Resonance
CSD	Content-Style Disentanglement
CVD	CardioVascular Diseases
C-S	Content-Style
CNNs	Convolutional Neural Networks
CONV	CONVolutional Layer
CHAOs	Combined (CT-MR) Healthy Abdominal Organ Segmentation
DC	Distance Correlation
DL	Deep Learning
DCT	Discrete Cosine Transformation
DRL	Disentangled Representation Learning
DREAM	Disentangled Representations for Efficient Algorithms for Medical data
ECG	ElectroCardioGram
ELBO	Evidence Lower-Bound Optimisation
FA	Factor-based Augmentation
FC	Fully-Connected
FID	Frechet Inception Distance
FiLM	Feature-wise Linear Modulation
GT	Ground Truth

GPU	Graphics Processing Unit
GRL	Gradient Reversal Layer
GANs	Generative Adversarial Networks
HD	Hausdorff Distance
HSIC	Hilbert-Schmidt Independence Criterion
IB	Information Bottleneck
IN	Instance Normalisation
ICA	Independent Component Analysis
IOB	Information Over Bias
i.i.d.	independent and identically distributed
I2I	Image-to-Image
KL	Kullback–Leibler
KLD	Kullback-Liebler Divergence
LR	Latent Regression
LV	Left Ventricle (or Ventricular)
LDDG	Linear-Dependency Domain Generalization
LPIPS	Learned Perceptual Image Patch Similarity
MI	Mutual Information
ML	Machine Learning
MR	Magnetic Resonance
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
MYO	MYOcardium
M&Ms	Multi-centre, Multi-vendor & Multi-disease Cardiac Image Segmentation
MOCO	MOmentum COntrast
MUNIT	Multimodal Unsupervised Image-to-Image Translation
MM-WHS	Multi-Modality Whole Heart Segmentation
MICCAI	International Conference on Medical Image Computing and Computer-Assisted Intervention
nnUNet	no-new UNet
NIFTI	Neuroimaging Informatics Technology Initiative
PANet	Pose Appearance Network

RA	Resolution Augmentation
RF	Radio Frequency
RV	Right Ventricle
ReLU	Rectified Linear Unit
RNNs	Recurrent Neural Networks
ResNet	Residual Network
S	Style
SIM	SIMilarity or Histogram Intersection
SVD	Singular Value Decomposition
SCGM	Spinal Cord Gray Matter Segmentation
SAML	Shape-Aware Meta-Learning
SOTA	State-Of-The-Art
SDNet	Spatial Decomposition Network
SPADE	SPAtially-adaptive DEnormalisation
STACOM	STatistical Atlases and COmputational Modeling of the Heart
TTT	Test-Time Training
VQ	Vector Quantisation
VAE	Variational Auto-Encoder
vMF	von-Mises-Fisher
XIL	Explanatory Interactive Learning

Chapter 1
Introduction

In the year 2018, Google initiated the deployment of an Artificial Intelligence (AI) program in Thailand with the aim of detecting and screening for Diabetic Retinopathy, a disease that leads to permanent blindness. The deep learning models trained with the high-quality patient data in Google's lab worked perfectly with more than 90% accuracy reaching a human specialist level. Millions of patients were about to be saved from the risk of permanent blindness with the fast and accurate diagnosis of deep learning models. However, the reality of the program was far from the expected results. Nurses were tasked with scanning a high volume of patients in a short span of time, often under challenging conditions such as poor lighting. As a result, a significant proportion of images was rejected by the deep learning models, leading to an increased need for manual review to ensure accuracy. In fact, nurses had to spend more time scanning each patient such that the imaging data matches the "taste" of AI. [1]

The differences between lab-based environments and reality, and the shifts between training data and test data, e.g. good and poor lighting conditions, do not only exist in Google's Thailand program. This issue is prevalent in nearly all tasks that utilise deep learning methods. Specifically, in medical image analysis, the significant statistical variations across data from various clinical centres (termed here *domains*) greatly affect the efficacy of deep models [5]. This reduction in performance is mainly caused by *domain shifts* due to differences in patient populations, scanners, and scanning acquisition settings, as discussed in [23]. Variations in patient populations can result in differences in underlying anatomy and pathology, owing to factors such as gender, age, and ethnicity, which vary across locations, as demonstrated in [24, 25, 26]. Additionally, variations in scanners and scanning acquisition settings can impact the characteristics of the acquired images, such as brightness and contrast [23].

The naive approach to handling domain shifts is to acquire and label as many and diverse data as possible. The acquisition of diverse data may be possible but privacy concerns have to be taken into account. In fact, there are many patients' data stored in safe havens e.g. within hospitals.

[1] Information in this paragraph is mainly from [22].

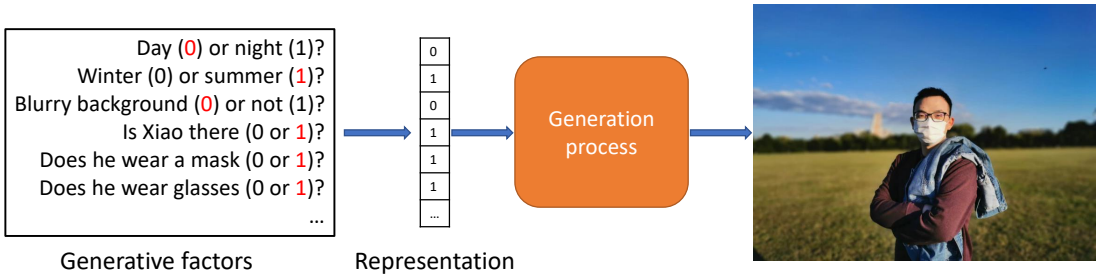


Figure 1.1: The illustration figure for demonstrating the generative factors, representation and generation process. The image is a profile photo of the author ("Xiao").

[2] However, annotating the data is considerably expensive and time-consuming. Taking the annotation of cardiac structure as an example, it takes several hours of work for a cardiovascular specialist [3] to label hundreds of 2-Dimensional Magnetic Resonance Imaging (MRI) images for one patient. Training a deep model for cardiac segmentation typically requires data from hundreds of patients for satisfactory performance. The cost of fully annotating the training data is remarkably high. Hence, taking advantage of unlabelled data plays a crucial role in modern deep learning techniques in medicine.

1.1 Motivation

This thesis focuses on tackling the domain shifts across domains as well as utilising unlabelled data in medical image analysis. An efficient and important approach is learning good representations that are invariant to the domain shifts without the requirement of extensive data annotations. Finding good representations for the task at hand is fundamental in machine learning and deep learning [28, 29]. In the context of representation learning, it is assumed that there is a generation process that produces the images with some generative factors. A simple example is illustrated in Fig. 1.1. The generative factors are encoded into a vector i.e. the representation. The generation process generates an image based on the representation. In representation learning with machine learning (ML) and deep learning (DL), typically some data are given and the target is to find the generation process and the generative factors. Under the modern DL framework, the generation process is modeled as neural networks [28]. The representations are modeled as vectors or tensors that are the features or the output of neural networks.

[2] A good example is that collaborating with Royal Infirmary, Edinburgh, we collected more than 4 Terabytes of patient data in our database.

[3] Average daily salary of cardiovascular specialists in the UK is around 189 GBP [27].

Disentanglement, a sub-area of representation learning, aims to separate out, or disentangle, the representations such that each representation approximates the underlying explanatory generative factor. A formal definition of disentangled representation is "*single latent units are sensitive to changes in single generative factors while being relatively invariant to changes in other factors*" [28]. As proposed in the famous Variational Auto-Encoder (VAE) [30], constraining each dimension of the latent vector representation to be independent surprisingly disentangles the latent space. Starting from VAE, enforcing the representations to be independent becomes a fundamental way to learn disentangled representations.

Disentanglement has great potential in addressing domain shifts towards more generalised deep models in the medical domain. Fundamentally, disentangling the factors that are consistent or invariant across domains and using these factors for the tasks at hand tackles the variation of data. For example, the MRI images of the same patient scanned in different hospitals contain the same anatomical information of the patient but with different appearances or intensities (modalities) caused by scanning with different scanners or the same scanner with different acquisition settings. Disentangling the anatomy factor from the modality factor and using the anatomy factor for downstream tasks produces better generalisation. More importantly, disentangled representations can be learnt with unlabelled data under limited supervision or with some expert knowledge [31]. This opens the door to taking advantage of the large amount of unlabelled data existing in the medical domain.

1.2 Challenges

It is evident that learning disentangled representations is well suited for addressing the domain shifts with the unlabelled data. In fact, some prior art [13] has already demonstrated such potential of disentangled representations. However, there are still several open problems and challenges in this area.

Although there are extensive works on disentanglement for computer vision tasks e.g. image-to-image translation [32, 33], facial attribute transfer [34, 35], pose estimation [36, 37]. There is no systematic study summarising the key theory in disentanglement and answering the questions – how to design disentanglement models; what are the measurements of disentanglement; and how the advances of disentanglement can be applied to medical data.

Moreover, many works have been proposed to evaluate the degree of disentanglement assum-

ing that the representations have the form of vectors. Recently, content-style disentanglement has been proposed to encode image "content" into a spatial tensor and image appearance or "style" into a vector, which has achieved state-of-the-art performance on many spatially equivariant tasks such as image-to-image translation. In the content-style disentanglement frameworks, different model designs, learning objectives, and data biases are employed for different computer vision tasks. This content-style disentanglement framework has also shown superior performance in the medical domain, as represented in [13]. While considerable effort has been made to measure disentanglement in vector representations, and assess its impact on task performance, such analysis for (spatial) content-style disentanglement is lacking.

Then, the benefit of leveraging disentanglement to design models that generalise well on new data has not been well studied. Generalisation on unseen data is the holy grail even in medical applications [38]. Although disentangled representations should be robust, recent studies [39, 40] found that disentanglement does not guarantee, for instance, combinatorial generalisation (understanding and producing novel combinations of familiar elements). More advanced approaches are required to learn generalisable disentangled representations with a guarantee.

Finally, the independence prior for disentanglement is too strong as an assumption that does not approximate well the true generative factors. When learning disentangled representation in real-life settings [41], statistical independence between latent variables does not hold when the generating factors are correlated [42, 43]. It is common that real data is not independent and identically distributed, and bias exists due to domain shifts. In these cases, it has been shown that factorisation-based inductive biases (as in VAE [30]) are not enough to learn the true generating factors. These biases can have significant implications for domain generalisation.

1.3 Overview and Technical Contributions

According to these problems, this thesis focuses on understanding the role of disentanglement in medical image analysis, measuring how different biases affect disentanglement and task performance, using disentangled representations to improve generalisation performance and exploring better representations beyond disentanglement.

More specifically, I first conduct a comprehensive survey on learning disentangled representations in the imaging domain. The theory of disentangled representation learning has been briefly concluded. Different frameworks enforcing disentanglement, the building blocks for

disentanglement as well as metrics of measuring disentanglement have been thoroughly revised. The survey is inspired by a tutorial I co-organised at the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) on disentangled representations (https://vios.science/tutorials/dream2021) and benefits greatly from feedback received from participants in the tutorial. Part of the content of the survey has been included in Chapter 3 and Chapter 7, where the publication is:

- **Liu, X.***, Sanchez, P.*, Thermos, S.*, O'Neil, A.Q. and Tsaftaris, S.A., 2022. Learning Disentangled Representations in the Imaging Domain. Medical Image Analysis, p.102516. *Equal contribution.

Chapter 3 is also benefited from:

- Fragemann, J., Ardizzone, L., **Liu, X.**, Tsaftaris, S.A., Egger, J. and Kleesiek, J., 2023. Review of Disentanglement Approaches for Medical Applications: Towards Solving the *Gordian Knot* of Generative Models in Healthcare. ACM Computing Surveys (under review).

The code repository for Chapter 3 is publicly available at https://github.com/vios-s/disentanglement_tutorial.

In Chapter 4, I examine the impact of various biases in the context of content-style disentanglement and determine the relationship between the degree of disentanglement and task performance. To achieve this objective, the following steps are taken: (i) A comprehensive analysis of the key design choices and learning constraints for three popular content-style disentanglement models is performed. (ii) The constraints of the models are relaxed or removed as ablations. (iii) The degree of disentanglement is measured using two proposed metrics, and its effect on task performance is evaluated. The results of the experiments indicate the existence of a "sweet spot" between disentanglement, task performance, and content interpretability. The findings show that an excessive emphasis on disentanglement may negatively affect model performance and the semanticness of content factors. The results of this study, together with the task-independent metrics used, provide valuable insights into the design and selection of models for applications that require disentangled content-style representations. Chapter 4 is based on the following publication:

- **Liu, X.***, Thermos, S.*, Valvano, G.*, Chartsias, A., O'Neil, A. and Tsaftaris, S.A., 2021. Measuring the Biases and Effectiveness of Content-Style Disentanglement. British Machine Vision Conference 2021. *Equal contribution.

The code for Chapter 4 is publicly available at https://github.com/vios-s/CSDisentanglement_Metrics_Library.

In Chapter 5, I propose methods to learn more generalisable disentangled representations. I have introduced two data augmentation techniques aimed at enhancing the domain adaptation and generalisation ability of state-of-the-art cardiac segmentation models. The "Resolution Augmentation" method creates a more diverse dataset by rescaling images to different resolutions within a range spanning different scanner protocols. The "Factor-based Augmentation" method projects the original samples onto disentangled latent spaces and combines anatomy and modality factors from different domains to generate more diverse data. However, these augmentations only produce more diverse training data and do not ensure that the disentangled representations will be able to generalise to unseen domains. Hence, it becomes necessary to consider more advanced approaches.

I have also investigated the use of meta-learning to improve the generalisation ability of disentangled representations. To this end, gradient-based meta-learning approaches, in which the training data are divided into meta-train and meta-test sets to simulate and tackle domain shifts during training, have demonstrated improved generalisation performance. However, the current fully supervised meta-learning approaches are not scalable for medical image segmentation as they require large efforts to create pixel-wise annotations. Furthermore, in low data regimes, the simulated domain shifts may not accurately represent the true domain shifts between source and unseen domains. To address these challenges, I propose a novel semi-supervised meta-learning framework with disentanglement. The framework explicitly models representations related to domain shifts, and disentangling and combining these representations to reconstruct the input image enables the use of unlabelled data to more accurately approximate the true domain shifts for meta-learning. As a result, the model can achieve better generalisation performance, especially when there is a limited amount of labelled data. The experiments have demonstrated the robustness of the proposed method on different segmentation tasks and have achieved state-of-the-art generalisation performance on two public benchmarks. Chapter 5 is based on the following publications:

- **Liu, X.**, Thermos, S., O'Neil, A. and Tsaftaris, S.A., 2021. Semi-supervised Meta-learning with Disentanglement for Domain-generalised Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2021.

- **Liu, X.**, Thermos, S., Chartsias, A., O'Neil, A. and Tsaftaris, S.A., 2020. Disentangled Representations for Domain-generalised Cardiac Segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 187-195). Springer, Cham.

Chapter 5 is also benefited from:

- Campello, V.M., ..., **Liu X.**, Tsaftaris, S.A., ..., Lekadir K. 2021. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. IEEE Transactions on Medical Imaging.

The code for Chapter 5 is publicly available at `https://github.com/vios-s/RA_FA_Cardiac` and `https://github.com/vios-s/DGNet`.

In Chapter 6, I consider compositionality as a prior to learning generalisable and interpretable representations. As discussed before, deep learning models often require a substantial amount of labelled data for effective training. On the other hand, humans are able to quickly identify crucial anatomy in medical images such as MRI scans with minimal instruction. This recognition capability is easily generalisable to new images from different medical facilities and to new tasks in different settings. This rapid and generalisable learning ability is largely due to the compositional structure of image patterns in the human brain, which is not well represented in current medical models. In Chapter 6, I examine the role of compositionality in learning more interpretable and generalisable representations for medical image segmentation. I propose that the underlying generative factors that produce medical images adhere to the compositional equivariance property, where each factor is both compositional (i.e. corresponds to structures in human anatomy) and equivariant to the task. As a result, a good representation that closely approximates the ground truth factor is compositionally equivariant. By modelling the compositional representations with learnable von-Mises-Fisher (vMF) kernels, I explore how different design and learning biases can be used to enforce compositional equivariance in un-, weakly-, and semi-supervised settings. In particular, for the semi-supervised setting, I evaluate the pro-

posed models on the task of semi-supervised domain-generalised medical image segmentation. The results show that our methods outperform several strong baselines. Chapter 6 is based on the following publications:

- **Liu, X.**, Sanchez, P., Thermos, S., O'Neil, A. and Tsaftaris, S.A., 2023. Compositionally Equivariant Representation Learning. IEEE Transactions on Medical Imaging (under review).
- **Liu, X.**, Thermos, S., Sanchez, P., O'Neil, A. and Tsaftaris, S.A., 2022. vMFNet: Compositionality Meets Domain-generalised Segmentation. In International Conference on Medical Image Computing and Computer Assisted Intervention 2022.

The code for Chapter 6 is publicly available at https://github.com/vios-s/vMFNet. Note that the paper "Compositionally Equivariant Representation Learning" has been submitted to the IEEE Transactions on Medical Imaging and is under review during the thesis writing.

1.4 Thesis Structure

Here I provide an overview of the thesis contents. Chapter 2 contains background information on medical imaging, as well as presents the datasets used. Chapter 3 presents a technical background on deep learning and representation learning, as well as a literature review on the main research areas of the thesis. Chapter 4 introduces our proposed metrics measuring content-style disentanglement. Chapter 5 introduces the solutions to generalisable disentangled representations. Then, Chapter 6 proposes new methods for compositional representation learning. Finally, Chapter 7 concludes the manuscript, discussing limitations and future extensions of this work.

Chapter 2
Clinical and Medical Imaging Background

In this chapter, I will give the clinical and medical imaging background that is relevant to the proposed methods in Chapter 4, Chapter 5 and Chapter 6. In this thesis, I mainly utilised magnetic resonance imaging (MRI) images. I first briefly demonstrate how MRI works. Cine-MRI is particularly discussed to motivate the segmentation task. Then, a discussion of the human heart and the spinal cord and gray matter is included. In particular, the anatomical variations and imaging variations causing domain shifts are presented. Finally, I discuss how the data preprocessing is performed on the multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset [5] and spinal cord gray matter segmentation (SCGM) dataset [7], which are heavily used in Chapter 5 and Chapter 6. For other datasets used in this thesis, the details and data preprocessing steps are discussed specifically in each chapter.

2.1 Magnetic Resonance Imaging

As a noninvasive imaging technology, MRI is becoming one of the gold standards for disease detection, diagnosis, and treatment monitoring [44]. In principle, MRI forms the visualisation of soft tissues by exciting and detecting the change in the direction of the rotational axis of hydrogen protons found in the water that makes up living tissues. As depicted in Fig. 2.1, an MRI scanner contains the major magnet, gradient coils and radio frequency (RF) coils. MRI scanning is performed with the following steps:

Alignment. Without any external magnetic field, the hydrogen protons spin in random directions in the human body. When the magnet of an MRI scanner applies a strong magnetic field B_0, the protons are aligned and spin parallel with or antiparallel to this external field B_0 at a frequency known as Larmor frequency.

Excitation. After aligning the hydrogen protons, the RF coil sends an electromagnetic RF

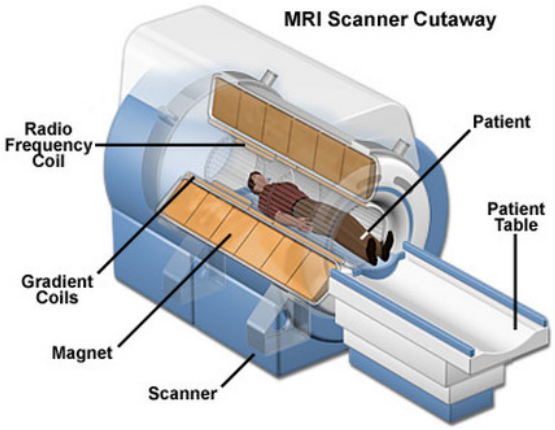


Figure 2.1: The illustration of the cutaway of an MRI scanner. Image is taken from [1].

pulse, which changes the spin direction of the hydrogen protons. This process is known as the excitation phrase.

Relaxation. After the RF pulse ends, the protons gradually change to the initial spin direction i.e. B_0 direction. During this relaxation phase, the protons release energy by emitting electromagnetic waves. The time required for the protons to reach 63% of the original spin direction is known as the 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 vary across different tissues due to variations in their water and fat content.

Detection. The emitted electromagnetic waves (the echo signal) can be detected by the radio frequency receiver coils in the MRI scanner. Note that the energy of the emitted electromagnetic waves decays over time. With the measured echo signal, the MRI scanner gathers the received information in the k-space. The final image in MRI is produced by applying the Fourier transform to the k-space data. K-space, also known as frequency space or spatial frequency domain, is a mathematical representation of the raw data acquired during an MRI scan. It is a 2D or 3D grid of points that contains information about the phase and spatial frequencies of the image pixels. I show an example of the frequency data and the corresponding image in Fig. 2.2.

Localisation. To localise the hydrogen protons that emit the signal, The gradient coils of the MRI scanner modulate the magnetic field B_0 in a predictable manner. This modulation causes a variation of the Larmor frequency of proton spins according to their position. Hence, the position is encoded in the echo signals enabling spatial localisation.

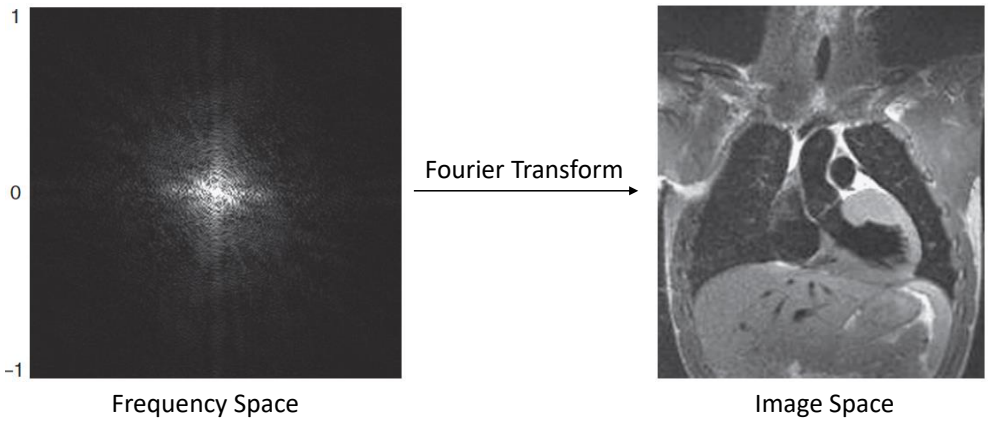


Figure 2.2: Reconstruction example from k-space (frequency space). Image is taken from [2].

There are many commonly used MRI protocols such as Brain MRI, Spine MRI and Cine-MR that contain standardised sets of procedures and parameters to acquire specific types of images. Different protocols provide significantly different information for visualising different aspects of tissues in the body, causing imaging variation across datasets. For example, T1-weighted and T2-weighted images are created by manipulating the timing of the signal acquisition to emphasise the differences in T1 and T2 relaxation times between tissues. For Brain MRI, T1-weighted images typically provide detailed anatomical information and T2-weighted images highlight pathology, such as edema and inflammation. In Section 2.1.1, I will describe more details about Cine-MR.

2.1.1 Cine-MR

Cardiovascular diseases (CVD) have been identified as the top cause of death globally by World Health Organization, accounting for 17.9 million lives each year [45]. Possible early signals of CVD include a raised blood pressure, level of glucose, and lipids, as well as obesity. Primary care facilities can easily measure these symptoms and if appropriate early treatment is provided for those at the highest risk of CVD, premature deaths can be efficiently prevented [46].

In clinics, cine-MR is the most commonly used protocol for CVD diagnosis, involving a temporal sequence with 10-30 frames. For cine-MR, the contraction and expansion of the heart are triggered by electrical signals that stimulate the myocardium, resulting in a consistently rhythmic cycle (as a heartbeat). This electrical activity is captured and measured using electrodes placed on the skin, as depicted in Fig.2.3, through a technique called electrocardiogram (ECG).

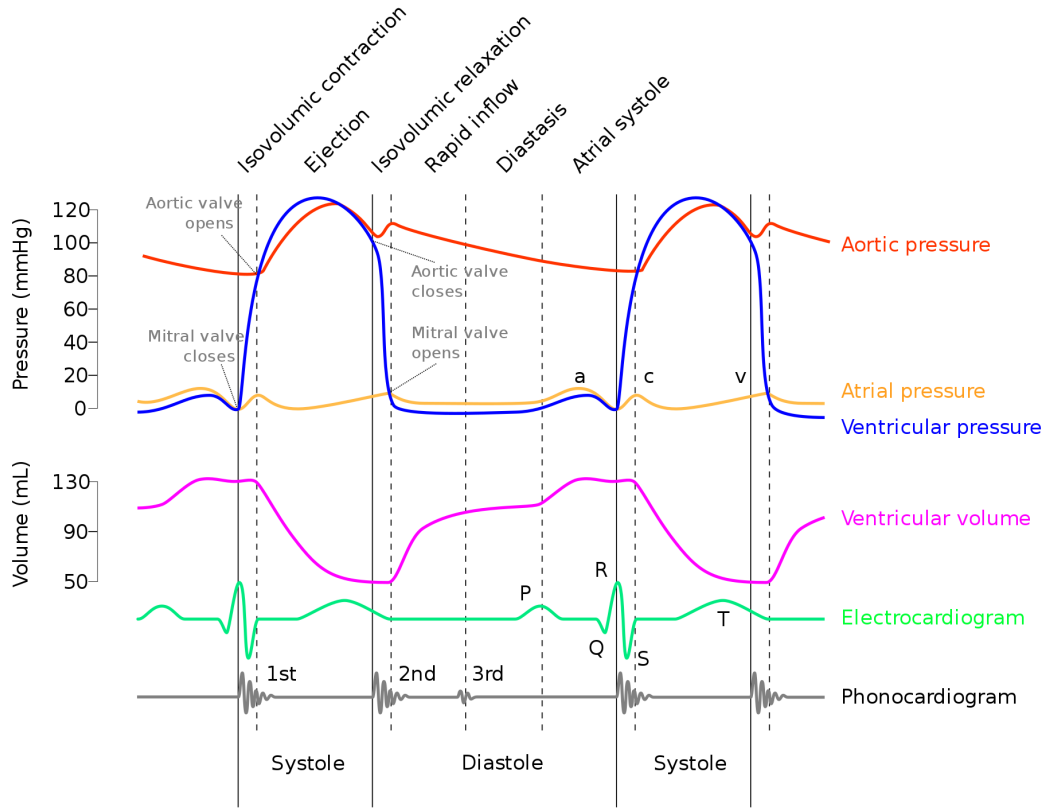


Figure 2.3: An example ECG showing the electrical activity of the heart, with the systole and diastole phases marked. Image is taken from [3].

ECG is used for imaging of the cardiac cycle. To ensure high-quality image acquisition, frequency space data for each frame are collected across different cycles. The synchronisation of the sampled data with specific frames is achieved using ECG gating, which detects the R-wave signal indicating the start of the systolic phase in the cardiac cycle. Both MR imaging and ECG pulse, which define the R-R interval of a heartbeat (see Fig. 2.3), are executed simultaneously, and synchronisation is carried out retrospectively. During an imaging session, multiple breath holds are required, and each cine-MR slice is scanned within approximately 10 seconds. To shorten the scanning time, non-isotropic images are acquired with a lower spatial resolution, typically with slice thickness ranging between $8mm$ and $10mm$.

Cine-MR is a bright-blood technique due to its high signal intensity within vessels compared to other tissues. It is frequently used to calculate functional indices like ejection fraction, wall thickness, myocardial mass and ventricle volumes for CVD diagnosis. These biomarkers are crucial for diagnosing and monitoring various cardiac conditions, including heart failure, coro-

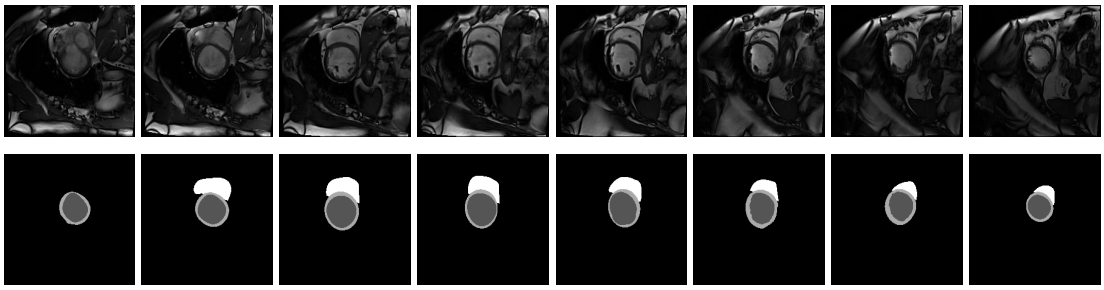


Figure 2.4: Example images and the corresponding segmentation masks of the end diastole frame of the cine-MR data. The segmentation masks mark the pixels of the left ventricle (circle shape), myocardium (torus shape) and right ventricle (white area). Images are taken from the M&Ms dataset [4]. Reproduced with permission.

nary artery disease, and congenital heart defects, etc. To obtain these functional indices, clinicians manually identify the different components of the heart (i.e. manual segmentation), such as the chambers, valves, and blood vessels, which is tedious and labour-intensive. Moreover, cross-corrections between clinicians are required to ensure accurate and consistent annotations of cardiac boundaries across all image slices and cardiac phases. In Fig. 2.4, I chose to show example cine-MRI images and the corresponding segmentation masks for the end diastole frame. This demand for accurate and automatic identification of the heart components motivates medical image segmentation with machine learning and deep learning techniques. Medical image segmentation refers to the process of separating an image into multiple segments or regions, each of which corresponds to a specific anatomical structure or tissue type, which is the major medical image analysis task considered in this thesis.

2.2 The Heart

In this thesis, the proposed methods in Chapter 5 and Chapter 6 were initially motivated by the clinical applications of cardiovascular disease analysis and diagnosis. Here, I specifically introduce the structure of the human heart and the mechanism of how the heart works. I also briefly discuss how the heart's anatomical structure varies across different populations. In particular, I discuss how aging affects the heart anatomical structure as an example of illustrating the variation of heart anatomy caused by different patient populations.

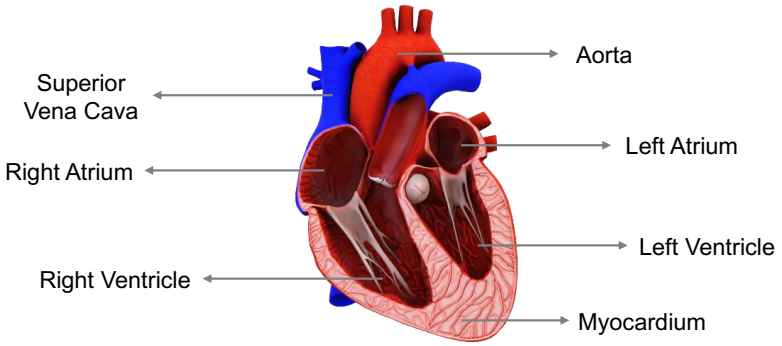


Figure 2.5: The cutaway figure of the human heart with labels to different anatomy. Image is produced based on the heart 3D model created by Microsoft Powerpoint.

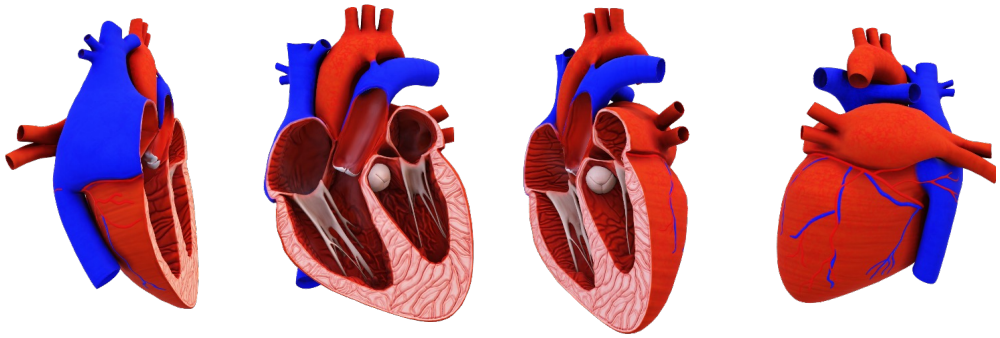


Figure 2.6: The cutaway figure of the human heart. 4 different views of the heart cutaway are depicted. Image is produced based on the heart 3D model created by Microsoft Powerpoint.

2.2.1 The structure of human heart

As illustrated in Fig. 2.5, the human heart contains four chambers i.e. the right atrium, right ventricle, left atrium and left ventricle. In Fig. 2.6, four different views of the cutaway of the human heart are depicted for better visualisation. Overall, the heart and lungs are the core organs of the human circulatory system. The heart and lungs oxygenate the blood before it is distributed to the rest of the body. The process of oxygenation occurs in several stages: **firstly**, deoxygenated blood from the body is received by the right atrium through the vena cava; **secondly**, the right atrium pumps the blood to the right ventricle; **thirdly**, the right ventricle pumps the low-oxygen blood to the lungs where it is replenished with oxygen; **fourthly**, oxygenated blood is received by the left atrium from the lungs, and then pumped to the left ventricle; **finally**, the left ventricle pumps the oxygen-rich blood through the aorta to the rest of the body. The whole process happens throughout a cardiac cycle from end systole to end diastole. At end

systole, the myocardium is fully contracted to pump blood. At end diastole, the myocardium is fully expanded to receive blood.

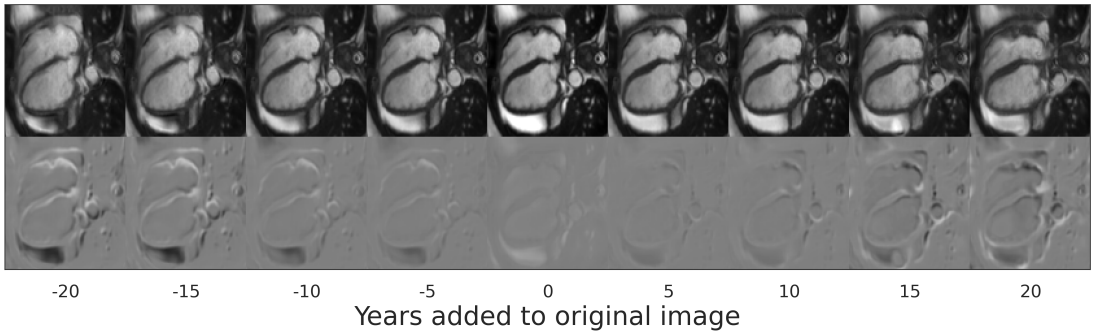


Figure 2.7: Synthetic aging cardiac MRI images. Image is taken from [4]. Reproduced with permission.

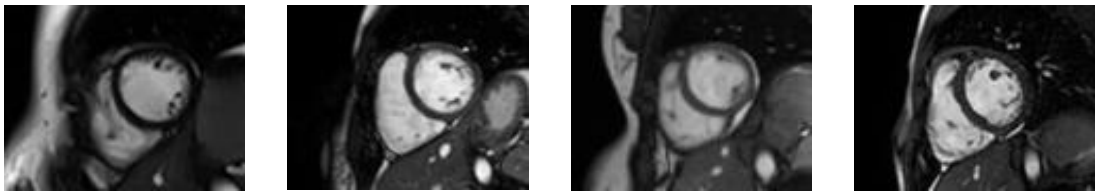
2.2.2 The anatomical variation

The differences in the anatomy of hearts across different clinical centres primarily arise from variations in patient populations. This variation in populations can affect the underlying anatomical and pathological features due to factors like age, gender, and ethnicity, which can vary among patients in different locations [24, 26]. For example, as I and the co-authors studied in [4], age has a positive correlation with morphological modifications in the heart, such as an enlarged left atrial diameter [47], increased thickness of the left ventricular (LV) wall, and reduced LV dimensions [48, 49]. These changes are linked with conditions like atrial fibrillation and heart failure with preserved ejection fraction [49, 50]. There are also gender-based differences in the aforementioned alterations, with women showing a higher prevalence of increased LV wall thickness [51]. Age-related deposition of epicardial adipose tissue has been also observed to increase significantly [51]. In Fig. 2.7, I show an example of synthetic aging cardiac MRI images to demonstrate the simulated anatomical variation caused by aging. The alterations tend to be spatially localised, with a focus on the interventricular septum and aorta, and these changes vary in opposite directions for different age ranges.

2.2.3 The imaging variation

Using different MRI scanners or under different acquisition settings in the same manner, the acquired MRI images can be significantly different. In Fig. 2.8, I chose to present 4 example

cine-MR images that are produced by four different scanners, which are Siemens MAGNETOM Avanto, Philips Achieva, General Electric Signa Excite and Canon Vantage Orian. Note that the magnetic field strengths of B_0 are the same for the four scanners i.e. 1.5 Teslas (1.5T). As Fig. 2.8 shows, the four images have significant differences in terms of contrast, brightness, resolution and noise levels, etc. Moreover, for different field strengths, the collected images will also vary in appearance. In principle, a stronger echo signal can be produced by the stronger field strength. Hence, a clearer image may be produced because the stronger signal overcomes more background noise. On the other hand, different scanners or different acquisition settings used may produce different artifacts in the images, resulting in a difference in image appearance. Overall, these factors cause domain shifts in the imaging characteristics across different clinical centres.



(a) Siemens (b) Philips (c) General Electric (d) Canon

Figure 2.8: Examples of cardiac MRI images that have anatomically similar structures and are produced by four different scanners. Images are taken from [5]. The figure is reproduced with permission.

2.3 Spinal Cord and Gray Matter

Apart from the cardiac data analysis, I also consider the other application of spinal cord and gray matter segmentation to verify the task robustness of the proposed approaches. For spinal cord and gray matter, I present the anatomical structure and functionality. As the population variation contributes little to the domain shifts, I chose to specifically demonstrate the domain shits caused by imaging characteristics variations. Finally, I discuss how the data preprocessing is performed on this dataset.

2.3.1 The structure of spinal cord and gray matter

The spinal cord is situated within the vertebral canal and facilitates the transmission of nerve impulses to 31 pairs of spinal nerves, enabling communication between the brain and periph-

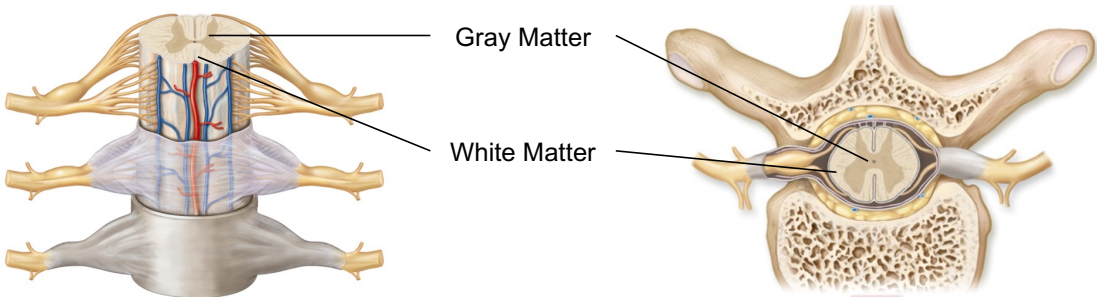


Figure 2.9: The illustration of the spinal cord with labels to the gray matter and white matter. Image is taken from [6].

eral nerves [52]. Two fundamental mechanisms underlie this transmission process, namely afferent signals that convey sensations originating from nervous tissue in the trunk, neck, and forelimbs to the brain, and efferent signals that transmit instructions from the brain to effector organs in the trunk, neck, and limbs, causing them to execute specific actions. In addition to these communication functions, the spinal cord is also responsible for regulating immediate and vegetative movements, such as reflex actions, central nervous system functions, and the sympathetic and parasympathetic systems. As depicted in Fig. 2.9, in a cross section of the spinal cord, the central region comprises the gray matter, which has a butterfly-like shape. Essentially, the gray matter is primarily made up of neuronal bodies and cells that modulate the immune system. The white matter contains axons that transmit information up and down the spinal cord. Segmenting the gray and white matter in the spinal cord plays a crucial role in the tissue specific analysis to help clinicians with diagnosis, prognosis, and treatment planning of diseases such as spinal cord injury, multiple sclerosis, and spinal cord tumours.

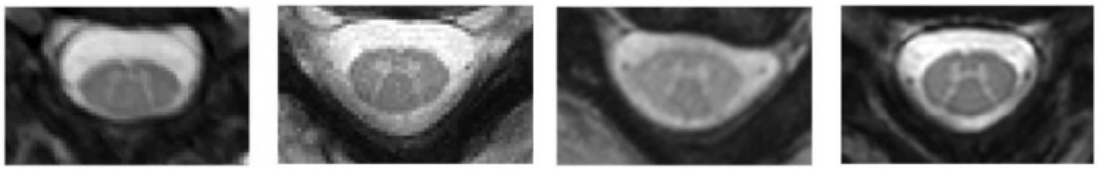


Figure 2.10: Examples of the MRI images for the spinal cord and gray matter segmentation. The images are scanned by different scanners on different sites i.e. UCL, Montreal, Zurich and Vanderbilt. Images are taken from [7].

2.3.2 The imaging variation

As shown in Fig. 2.10, I present 4 example images that are scanned in different sites with different scanners. Although the variation in the anatomical structure is negligible, the variation in image appearance introduces remarkable domain shifts. Similar to cine-MR data, different images from different sites have varying contrasts, brightness, resolutions and noise levels, etc.

2.4 Data Preprocessing

In this thesis, I mainly used the multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset [5]. The dataset contains 320 subjects. Subjects were scanned at 6 clinical centres in 3 different countries (Spain, Germany and Canada) using 4 different magnetic resonance scanner vendors (Siemens, Philips, General Electric, and Canon). For each subject, only the end systole and end diastole phases are annotated. Voxel resolutions range from $0.85 \times 0.85 \times 10$ mm to $1.45 \times 1.45 \times 9.9$ mm. The number of time frames for the subjects ranges from 25 to 30. For more details about this dataset, I refer the readers to the M&Ms challenge page (https://www.ub.edu/mnms/) and the accompanying journal paper [5]. For data preprocessing, I first split the data into 4 subsets based on the scanner vendor. In Chapter 5 and Chapter 6, each subset is considered as one domain for the task of domain generalisation. The 4 subsets or domains have 95, 125, 50 and 50 subjects. For each subject, the data is in Neuroimaging Informatics Technology Initiative (NIFTI) format [53], which has 4 dimensions corresponding to the number of slices for each frame, the number of time frames, the height of the image and the width of the image. For each subject, the 4-Dimensional data is re-stored as hundreds of 2-Dimensional (2D) images. For each 2D image, multiple augmentation techniques are applied including random rotation, and random scaling with a scaling ratio ranging from 0.8 to 1.2, random cropping to the size of 288×288, random horizontal and vertical flipping and adding Gaussian noise with the kernel size 5 and the randomly chosen deviation in the range of 0.25 to 1.25. All the augmentations are performed with the embedded functions in PyTorch [54]. The data preprocessing code is publicly available at https://github.com/vios-s/DGNet.

The other dataset, spinal cord and gray matter segmentation (SCGM) dataset, is heavily used in Chapter 5 and Chapter 6. The SCGM dataset is collected from 4 different medical centres in UCL, Montreal, Zurich and Vanderbilt with 3 different MRI scanners (Philips Achieva, Siemens

Trio, Siemens Skyra), which are considered as 4 domains in this thesis. The voxel resolutions range from $0.25 \times 0.25 \times 2.5$ mm to $0.5 \times 0.5 \times 5$ mm. Each domain has 10 labelled subjects and 10 unlabelled subjects. For more details of the SCGM dataset, I refer the readers to the dataset page (http://niftyweb.cs.ucl.ac.uk/challenge/index.php) and the accompanying journal paper [7]. For data preprocessing, the 2D images are randomly cropped into the size of 144×144. The data preprocessing code is publicly available at https://github.com/vios-s/DGNet.

2.5 Summary

In this chapter, I described the necessary clinical and medical imaging background for understanding the main technical contributions and clinical impact of the proposed methods. In particular, MRI and the cine-MR protocol are briefly introduced. Then, the anatomical structure, the anatomical and imaging variations of the heart and the spinal cord and gray matter are presented. Finally, the data preprocessing is described for the datasets used throughout the thesis. In the next chapter, I will introduce the technical background covering representation learning and compositionality.

Chapter 3
Technical Background

In this chapter, I will give the technical background that is relevant to the proposed methods in Chapter 4, Chapter 5 and Chapter 6. To define disentanglement I first revisit key concepts in learning representations. I then provide an overview of key generative frameworks forming the basis of many subsequent models; building blocks of disentanglement; and evaluation metrics. Then, I discuss the lessons that we can learn from computer vision tasks for representation learning in medical image analysis. Last but not least, I briefly review compositionality in deep learning for the compositional representation learning that is studied in Chapter 6. This chapter is also accompanied by a repository offering links to the implementations of key methods and to existing metrics: https://github.com/vios-s/disentanglement_tutorial. For a more detailed discussion of disentangled representation learning (DRL) theory and the applications of DRL on medical tasks, I refer readers to our surveys [12] and [55]. The tutorial (DREAM 2021: Disentangled Representations for Efficient Algorithms for Medical data https://vios.science/tutorials/dream2021) and workshop (Medical Applications with Disentanglements https://mad.ikim.nrw/) I co-organised at the International Conference on Medical Image Computing and Computer Assisted Intervention also provide useful information.

3.1 Introduction

Imagine the need to develop a method to localise the ventricles in Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) scans of the brain in patients. This method must be robust to any changes in the imaging process, scanner, and noise, as well as to anatomical and pathological variation. The current deep (supervised) learning paradigm indicates that we *must*

This chapter is based on:

- **Liu, X.***, Sanchez, P.*, Thermos, S.*, O'Neil, A.Q. and Tsaftaris, S.A., 2022. Learning Disentangled Representations in the Imaging Domain. Medical Image Analysis, p.102516. *Equal contribution.

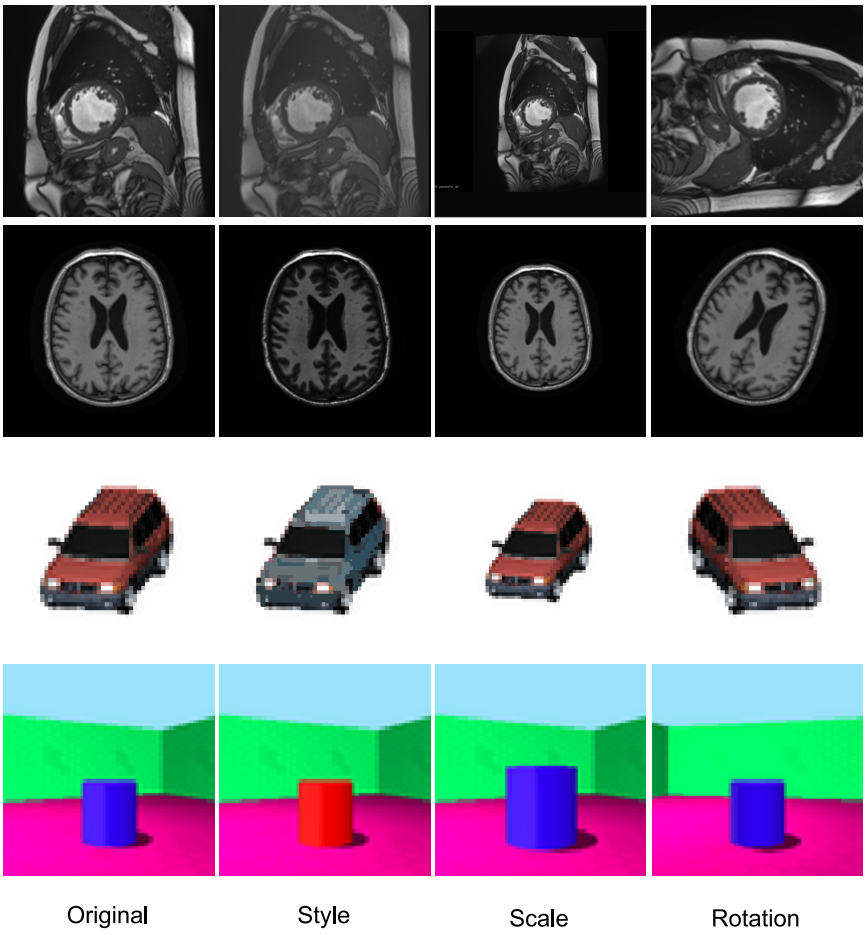


Figure 3.1: Examples of factors of variations: style, scale, and rotation in the context of cardiac scans [8], brain scans [9], cars [10], and 3D shapes [11]. This figure was originally created by Pedro Sanchez. Reproduced with permission.

present to the system as many examples as possible to instill robustness by learning what is unnecessary, or nuisance [56], e.g. the patient being placed at a rotated angle in the scanner, as opposed to what matters, i.e. the location of the ventricle. However, collecting and annotating enough data to cover such real-world variation is an unrealistically time-consuming and costly solution.

Surprisingly, we may not always need annotated data or carefully crafted data augmentations to achieve this. With DRL, one learns to encode the underlying factors of variation into separate latent variables [28, 57], which ultimately capture sensitive and useful information for the task at hand and also understand the underlying causal relations amongst the variables. I choose to introduce the reader to DRL by presenting 3 indicative examples of disentangled factors in

Fig. 3.1, which affect the colour, scale, and rotation of the rendered object in the corresponding scene. By adopting DRL, one can design deep models that will be robust to representations from unseen domains, a result that cannot always be achieved through data augmentation.

3.2 Key Concepts in Representation Learning

Notation. I use x, $\mathbf{x}$ and $\mathbf{X}$ to denote scalars, vectors, and higher-dimensional tensors respectively, drawn from the domain $\mathcal{X}$ of corresponding dimensions. I use $\mathbf{X}_i$ to refer to a datum of the above tensors (of any dimension) for presentation simplicity where tensor dimensionality is implied by the context. I will assume we have access to a dataset containing samples of $\mathbf{X}_i$, where $i \in [1, N]$, N denoting the number of samples. I use $\mathcal{X}$ to denote the observed variables of the input domain, $\mathcal{Z}$ for latent representations, $\mathcal{S}$ for real generating factors, and $\mathcal{Y}$ for the output domain. For example, if we choose to solve a classification task, then $\mathcal{Y}$ is a space of scalars y.

3.2.1 Model learning

Considering the task of learning a mapping between two domains [58] i.e. $f : \mathcal{X} \to \mathcal{Y}$, one can split f into two components, $f : \boldsymbol{E}_\phi \circ \boldsymbol{D}_\theta$. $\boldsymbol{E}_\phi$ maps to an intermediate latent representation $\mathcal{Z}$ ($E_\phi : \mathcal{X} \to \mathcal{Z}$) whereas $\boldsymbol{D}_\theta$ maps to the output ($\boldsymbol{D}_\theta : \mathcal{Z} \to \mathcal{Y}$). I will term E_ϕ the "encoder" and $\boldsymbol{D}_\theta$ the "decoder".[1] Thus, the goal of model learning is the solution to the task at hand by learning a good representation. Below, I discuss the desirable properties of a good representation.

3.2.2 Representation learning

Finding good representations for the task at hand is fundamental in machine learning [28, 29]. Consider the task of detecting brain tumours by placing a bounding box $\mathbf{Y}_i$ around each tumour in the image $\mathbf{X}_i$. A dataset may contain brain samples with different morphologies, acquired using different protocols in different sites (hospitals), etc. The goal is to create a representation suitable for the task. If the tumour changes location in the image, we would like the bounding box output to change location accordingly; the representation will be *equivariant* to the location

[1] $\boldsymbol{D}_\theta$ is often referred to as a classifier or a regressor, however, I avoid this nomenclature here to be more general.

of the object of interest. On the other hand, we would like the representation to be *invariant* to acquisition-related changes.

Symmetries. Symmetries Ω are *transformations* that leave some aspects of the input intact [59, 60, 61]. For instance, the category of an object does not change after applying shift operations to the image, therefore these operations are considered symmetries in the object recognition domain. Using the model f and symmetries Ω, I now proceed to define the equivariance and invariance properties.

Equivariance. A mapping $\boldsymbol{E}_\phi : \mathcal{X} \to \mathcal{Z}$ is equivariant w.r.t. the group Ω, if there is an action (transformation in our case) of the group $\omega \in \Omega$ of the input $\mathbf{X} \in \mathcal{X}$ that affects the output $Z \in \mathcal{Z}$ in the same manner. Formally, this means that Ω-equivariance of $\boldsymbol{E}_\phi$ is obtained when there exist mappings $M_\omega : R^d \to R^d$ and $M'_\omega : R^{d'} \to R^{d'}$ applying ω to the input and the output such that:

$$\boldsymbol{E}_\phi(M_\omega \circ \mathbf{X}) = M'_\omega \circ \boldsymbol{E}_\phi(\mathbf{X}),\ \forall \omega \in \Omega. \tag{3.1}$$

In practice, one chooses transformations that induce the desired equivariance and learned properties in accordance with the task at hand, thus a good understanding of the problem (also known as *domain knowledge*) is required [62]. Classical examples where equivariance to translation, shift, and mirroring might be important, are image segmentation, pose estimation, and landmark detection tasks. Note that the compositional equivalence theory proposed in Chapter 6 is based on the definition here. I will revise the definition of equivariance in Chapter 6.

Invariance. Formally, $\boldsymbol{E}_\phi$ is invariant to transformations of Ω if:

$$\boldsymbol{E}_\phi(M_\omega \circ \mathbf{X}) = \boldsymbol{E}_\phi(\mathbf{X}),\ \forall \omega \in \Omega. \tag{3.2}$$

The transforms we want to adhere to are usually task-specific. In Chapter 4, I will introduce the transformations used in popular content-style disentanglement methods.

3.2.3 Generating factors

Considering a distribution that characterises the domain $\mathcal{X}$, the *generating factors* $\mathcal{S}$ are the underlying variables that fully characterise the variation of the data –seen or expected to be seen. Recent studies [28, 29] argue that representations should enable the decomposition (i.e. disentanglement) of the input data into separate factors. Each factor should correspond to a variable

of interest in the underlying process that generated the data. For the rest of the chapter, I will refer to the real-world generating factors as "real" and to those learned by a model as "learned". [2] In the brain tumour detection example, several variables such as tumour texture/location, brain shape, acquisition protocol, image contrast, etc. may be involved. In general, the more complex the image, the more variables, and the higher the number of possible combinations. Enumerating all these combinations readily leads to a combinatorial explosion in the possible combinations that a dataset must contain to enable a model to learn (from data alone) the desired in/equi-variances. It is not realistic to identify every factor and cover every possible combination. Domain knowledge enables the elucidation of as many factors as possible and allows us to define which real factors we want to be in/equi-variant to.

3.2.4 Domain shifts

An i.i.d. data distribution is easy to consider but forms a strong and often unrealistic assumption. All non-synthetic datasets are somewhat biased due to the finite nature of the acquired data. If learning algorithms are trained with standard supervised learning [58] without additional assumptions, there is little hope that the learned function will be robust to domain shifts. A model's ability to maintain the desired behaviour across domain changes is also referred to as *out-of-distribution* generalisation [63]. For the brain tumour detection example, both CT or MRI scanners acquire images, but we might know that a given hospital uses CT. In this case, modality-related factors are linked to the hospital-related variables. Therefore, understanding the data generation process and the underlying relations between variables can help to distill the important visual information, and to create mechanisms that are more generalisable. Such reasoning enables the design of principled strategies for mitigating the data bias [64]. In fact, we can explicitly define the changes we want our model to be invariant or equivariant to, by modeling domain shifts such as: i) population i.e. different cohorts, ii) acquisition i.e. different cameras, sites or scanners, and iii) annotation shift i.e. different annotators.

3.2.5 Disentangled representations

Disentangled representations can address some of the challenges described until now by learning representations with equi/in-variances to specific undesired variables, whilst considering the

[2] In Chapter 6, I will define that the generative factors satisfy the compositional equivariance property.

data generation process and potential domain shifts. Although a widely accepted definition of disentangled representations is yet to be defined, the main intuition is that by disentangling, we separate out the main factors of variation that are present in our data distribution [28, 57, 65, 31]. I characterise a factor as "disentangled" when any intervention on this factor results in a specific change in the generated data [65, 66].

3.2.5.1 Formalising disentanglement

Higgins et al. [57] have recently presented a generic definition for disentanglement. Given a compositional world W and a set of transformations Ω (as defined in Section 3.2.2), they define a function $f : W \rightarrow Z$ that can induce Ω in the latent representation $Z \in \mathcal{Z}$ in an equivariant manner. The representation Z is defined as "disentangled" if there is a decomposition $Z = Z_1 \times \cdots \times Z_n$ such that a transformation ω applied on Z_i will result in an equivalent transformation in the input domain $\mathcal{X}$, leaving all other aspects controlled by $Z_{j \neq i}$ unchanged. This definition meets the desired properties of a disentangled representation as defined by several works in DRL [28, 67, 18, 68]: a) modularity i.e. each latent dimension should encode no more than one generative factor, and b) informativeness i.e. all underlying generative factors are encoded in the representation.

A complementary view to the definition of Higgins et al. [57] comes from the Information Bottleneck (IB) principle introduced in [69]. IB allows for learning "good" representations for the task at hand, by trading-off sufficiency and complexity. Adopting IB, Achille et al. [56] argue that such representations should be: i) *sufficient* for the task, meaning that we do not discard information required for the output; ii) among all sufficient representations, it should be *minimal* retaining as little information about the input as possible; and finally iii) it should be *invariant* to nuisance effects so that the final classifier will not overfit to any correlations between the dataset nuisances and the ground truth labels.

3.2.5.2 Identifiability

Learning disentangled representations without any type of supervision is impossible as an infinite family of models that could have generated the observed data exist [31]. Thus, *identifying* the model that generated the data without any additional information is impossible. Given an observation $\mathbf{X}_i$, there is an infinite number of generative models that could have generated a

sample from the same marginal distribution [31, 70, 71].

This follows from prior work in non-linear independent component analysis (ICA) [72]: even though the linear case is identifiable, the flexibility given by the non-linear case makes it non-identifiable without extra information. Recently Khemakhem et al. [71] bridged the gap between non-linear ICA and other deep latent variable models, and showed that unsupervised disentanglement methods are non-identifiable without additional assumptions.

3.2.5.3 Disentanglement as inductive bias

The solution to identifiability is the use of domain knowledge i.e. the *inductive bias*, instead of using explicit supervision [31, 70, 71]. Current representation learning already benefits from the inductive biases of Convolutional Neural Networks (CNNs) [73] and Recurrent Neural Networks (RNNs) [74]. Outside of the visual domain, language has been modeled with recurrent neural networks that capture the sequential nature of data for making predictions [75]. Recent attention and self-attention models, such as the transformer architecture [76], focus on learning the internal structure of the input data. These self-attention models essentially approximate the best inductive biases for each sample in the data distribution. Overall, disentanglement priors add structure to the learned representations to better correspond to the underlying generation process. It is this useful bias that makes the utilised models identifiable. One of the goals of this chapter is to highlight the various inductive biases used.

3.3 Frameworks Enforcing Disentanglement

3.3.1 Variational autoencoders

Auto-Encoders (AEs) or Variational Auto-Encoders (VAEs) [30, 77] decompose factors via image reconstruction [78, 79]. A typical VAE, depicted in Fig. 3.2(a), discovers and disentangles factors of variation by forcing independence between different dimensions of $\mathbf{z}$, while reconstructing the input $\mathbf{X}$. Notably, the three content-style disentanglement methods I study in Chapter 4 are either based on VAEs or inspired by VAEs.

A widely-used VAE that encourages disentanglement is the β-VAE [80]. Its main objective is

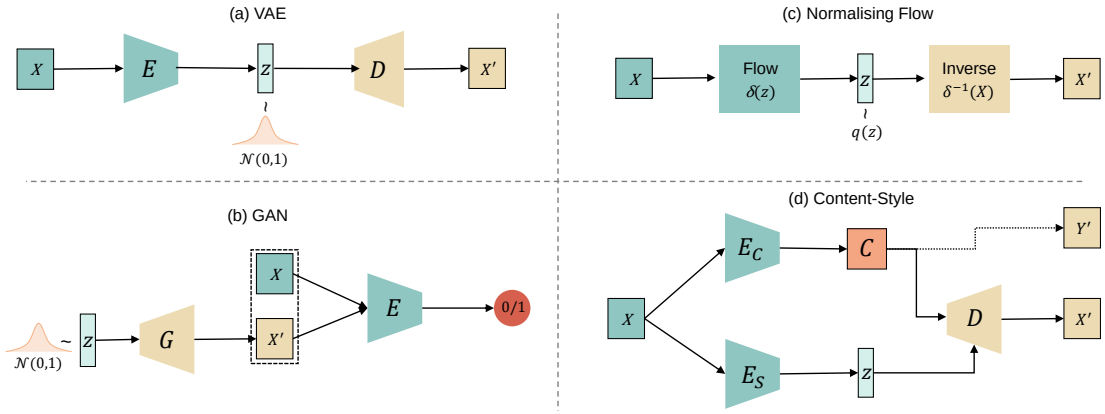


Figure 3.2: Fundamental architectures for disentanglement: a) VAE, b) GAN, c) Normalising Flows, d) Content-Style disentanglement. $\mathbf{X}$ and $\mathbf{X}'$ are the input and reconstructed images. $\mathbf{z}$, $\mathbf{C}$ are the latent representations, where $\mathbf{C}$ represents a tensor latent variable (e.g. image content) and $\mathbf{z}$ represents a vector latent variable. The dashed line in (d) denotes the use of $\mathbf{C}$ for learning a representation $\mathbf{Y}'$ for a parallel equivariant task (e.g. semantic segmentation). Finally, $\mathcal{N}$ denotes the normal distribution with zero mean and unit variance, whilst $q(\mathbf{z})$ can be any prior distribution. This figure is taken from [12]

the maximisation of the Evidence Lower-Bound Optimisation (ELBO):

$$\mathcal{L}_{ELBO}(\theta, \phi; \mathbf{X}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}[\log p_\theta(\mathbf{X} \mid \mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{X})||p_\mathbf{z}), \tag{3.3}$$

to balance (via $\beta > 1$) the reconstruction error versus adherence to the approximate posterior $q_{\mathbf{z}|\mathbf{X}}$ from the latent prior $p_\mathbf{z}$. $p_\theta(\mathbf{X} \mid \mathbf{z})$ is modeled as the decoder with weights θ. $q_\phi(\mathbf{z} \mid \mathbf{X})$ is modelled as the encoder with weights ϕ. The first term can be expanded as:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}[\log p_\theta(\mathbf{X} \mid \mathbf{z})] = \int_\mathbf{z} q_\phi(\mathbf{z} \mid \mathbf{X}) \log p_\theta(\mathbf{X} \mid \mathbf{z}) \tag{3.4}$$

The Kullback–Leibler divergence is defined as:

$$D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{X})||p_\mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X})}[\log q_\phi(\mathbf{z} \mid \mathbf{X}) - \log p_\mathbf{z}] = \int_\mathbf{z} q_\phi(\mathbf{z} \mid \mathbf{X}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{X})}{p_\mathbf{z}}. \tag{3.5}$$

Note that $p_\mathbf{z}$ is usually a normal distribution with identity covariance matrix $\mathcal{N}(0, \mathbf{I})$. The diagonal covariance forces an orthogonal factorisation of the latent space, similar to a principle

component analysis, which reasonably explains the disentanglement capabilities of VAEs [81, 82]. A $\beta > 1$ encourages disentanglement by forcing $q(\mathbf{z} \mid \mathbf{X})$ to carry less information about the reconstruction by increasing the weight of the Kullback–Leibler divergence term [82] and consequently, increasing independence between the factors of $\mathbf{z}$.

3.3.2 Generative adversarial networks

Generative Adversarial Networks (GANs) [83], see Fig. 3.2(b), typically employ a generator $\boldsymbol{G}$ and a discriminator $\boldsymbol{D}$ in an adversarial game. $\boldsymbol{G}$ generates an image by sampling from an isotropic Gaussian distribution, while $\boldsymbol{D}$ is given the synthetic image and a real one ($\mathbf{X}$), and tries to identify which input is real/fake. The game is formalised as:

$$\min_{\boldsymbol{G}} \max_{\boldsymbol{D}} \mathcal{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}) = \mathbb{E}_{p(\mathbf{z})}[\log(1 - \boldsymbol{D}(\boldsymbol{G}(\mathbf{z}))] + \mathbb{E}_{p(\mathbf{X})}[\log(\boldsymbol{D}(\mathbf{X}))], \tag{3.6}$$

where $\mathbf{z}$ is a vector with values sampled from the aforementioned Gaussian and $\boldsymbol{G}(\mathbf{z})$ is the generated image. Recent advances in GAN design and training have led to high-fidelity image generation [84, 85, 86]. GANs can learn disentangled representations by adding regularisation terms [67], by creating an architectural prior [84], or even by a post-hoc decomposition of the learned manifold after training [87].

3.3.3 Content-style disentanglement

The aforementioned models typically decompose factors into a single vector representation. However, a recent trend in disentanglement focuses on the decomposition of the input image into different latent variables that encode different properties, such as geometry vs. style. This form of disentanglement is the so-called Content-Style Disentanglement (CSD) [88], where an image is decomposed into domain-invariant "content" and domain-specific "style" representations [89, 90]. Most works in CSD encode content in spatial (tensor) representations to preserve the spatial correlations and exploit them for a spatially equivariant task, such as Image-to-Image (I2I) translation [32, 33] and semantic segmentation [13]. The corresponding style i.e. the information that controls the image appearance such as colour and intensity, is encoded in a vector. An abstract visualisation of a CSD model is depicted in Fig. 3.2(d). Note that decomposing content from style is not a trivial process, and encoding content as a high-dimensional representation is not enough. Recent work introduces several design (in terms of the model

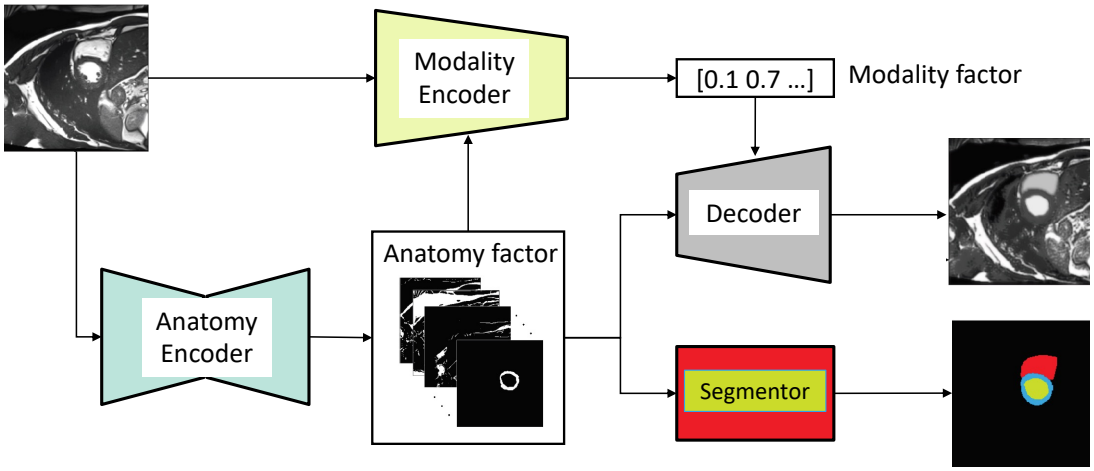


Figure 3.3: The Spatial Decomposition Network (SDNet). Two paths are designed in the model i.e. reconstruction and segmentation. The input image is decomposed into a spatial anatomy space and a vector modality space. Combining the two factors reconstruct the image. The segmentation mask is predicted with the anatomy factor as input. Figure is reproduced with permission of Chartsias et al. [13].

architecture) and learning (in terms of loss functions) biases to achieve this separation. I denote these inductive biases as "building blocks" and discuss them in the following section.

3.3.3.1 SDNet

A frequently referred content-style disentanglement framework in this thesis is Spatial Decomposition Network (SDNet) [13]. Here, I discuss SDNet thoroughly. SDNet decomposes 2D medical images into spatial anatomical factors (content) and non-spatial modality factors (style). Regarding single-modal medical image segmentation, the input images are acquired with only one modality e.g. MRI images. When temporal information is available, temporal consistency objectives can be applied to boost the performance as in [91]. Based on SDNet, Jiang et al. [92] additionally disentangle the pathology factor to perform semi-supervised pathology segmentation. SDNet-based methods in segmentation also provide the possibility to handle the domain shifts across different domains Additionally, the variational encoding of the style representation allows for sampling and interpolation of the appearance factors, enabling the synthesis of new plausible images [93]. To learn generalisable representations, gradient-based meta-learning can be applied as a learning strategy when giving multi-domain data [16]. Shin et al. [94] disentangle intensity and non-intensity for domain adaptation in CT images.

Kalkhof et al. [95] also disentangles content from style information using a conditional GAN for cross-domain segmentation.

As shown in Fig 3.3, SDNet uses two different encoders for factorising content into a spatial representation and style into a vector one. A decoder is responsible for reconstructing the input by combining the two latent variables, while a segmentation module is applied on the content latent space to learn to predict the segmentation mask for each cardiac part. SDNet learns the content which is represented as multi-channel binary maps of the same resolution as the input. This is obtained with a softmax and a thresholding function. To encourage the style encoder to encode only style-related information, the authors employ a VAE network. Then, style and content are combined to reconstruct the input image by applying a series of convolutional layers with FiLM layers [96] (see Section 3.4). SDNet has been extensively evaluated on the ACDC [8], MM-WHS [97, 98, 99], CHAOs [100], and M&Ms [5] cardiac datasets, as well as on the SCGM [7] spinal one.

3.4 Disentanglement Building Blocks

I now describe common layers and modules that are used at various levels of the model design to encourage disentanglement. I associate these so-called building blocks with different high-level parts of the aforementioned AEs and generative models. Note that typically several of these are combined. In principle we would like to have the minimal set required to solve the task, noting that at times these blocks can compete.

3.4.1 Encoding modules

The following are commonly used at various levels of the encoder(s) in popular architectures as bottlenecks. I use representation bottlenecks as a way of reducing the amount of information in the data which will force the network to encode mainly useful concepts.

Instance normalisation. Instance Normalisation (IN), originally proposed in [101] for style removal, is commonly used after each convolutional layer of the content encoder to suppress style-related information. In fact, IN removes any contrast-related information from each instance (data sample), encouraging content-related features to be propagated to the following layers. An indicative example is the content encoder in [102], where IN replaces all batch

normalisation layers [103].

Average pooling. Contrary to IN, average–pooling or global–pooling is commonly used to suppress the content information in the style encoder [102]. By averaging values and flattening a spatial feature into a vector, this operator removes any spatial correlation and encodes the global mean statistics (i.e. image style).

Parsimony. For CSD models that require semantic and parsimonious content for parallel spatially equivariant tasks, there is a need for discretisation of the encoded continuous information. Such discretisation also can help to remove style-related information. The Gumbel Softmax operator is a differentiable solution to this problem. This operator mimics the reparametrisation trick performed in VAEs by sampling from a standard Gumbel distribution and using the Softmax as an approximation of the "argmax" step that is usually coupled with one-hot operators for discretisation. Another tool that can further restrict the amount of information in a latent space is known as Vector Quantisation (VQ) [104]. VQ uses a dictionary of learnable entries to restrict the latent features to discrete set of values.

3.4.2 Entanglement modules

Effective recombination or entanglement of the content and style representations in a decoder is vital. The following approaches or layers are commonly used for this purpose at various levels of the decoder in popular CSD architectures.

Concatenation. Simple concatenation allows the content and style to be more flexible in capturing the desired information [33, 37]. However, this may limit the controllability of learning the content and style as the representations may not capture desired information e.g. style representation capturing the shape information.

Adaptive instance normalisation. The Adaptive Instance Normalisation (AdaIN) layer [102] is commonly used at multiple decoder levels to recombine the content and style representations. As depicted in Fig. 3.4(a), each AdaIN layer performs the following operation:

$$\text{AdaIN} = \gamma \frac{C_j - \mu(C_j)}{\sigma(C_j)} + \beta_j, \tag{3.7}$$

where each feature map C_j is first normalised separately by subtracting the mean $\mu(C_j)$ and dividing the variance $\sigma(C_j)$ of the feature map, and then is scaled and shifted based on γ and

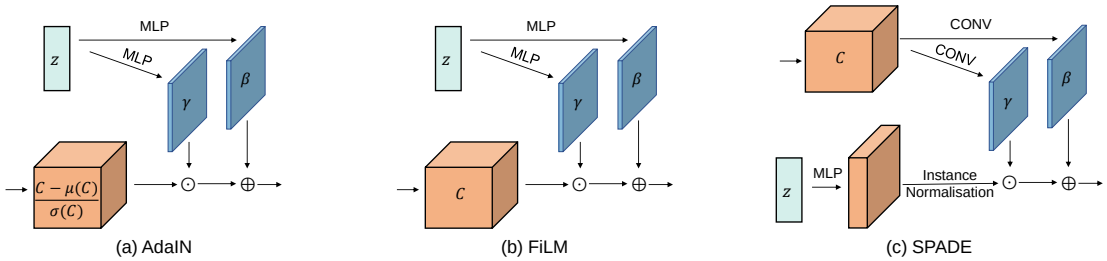


Figure 3.4: Disentanglement building blocks that combine content $\mathbf{C}$ with style $\mathbf{z}$: a) AdaIN, b) FiLM, and c) SPADE. $\odot$ and $\bigoplus$ denote element-wise multiplication and addition, respectively. MLP and CONV denote multilayer perceptron and convolutional layers. This figure is taken from [12].

β, which are parameters of an affine transformation of the style representation (adaptive mean and standard deviation).

Feature-wise linear modulation. As shown in Fig. 3.4(b), Feature-wise Linear Modulation (FiLM) [96] is similar to AdaIN. FiLM was initially proposed as a conditioning method for visual reasoning (the task of answering image-related questions). Using FiLM, each channel of the network's intermediate features C_j is modulated based on γ_j and β_j as follows:

$$FiLM(C_j|\gamma_j, \beta_j) = \gamma_j \cdot C_j + \beta_j, \tag{3.8}$$

where element-wise multiplication $(\cdot)$ and addition are both broadcast over the spatial dimensions. It is used in [13] to combine the content and style in the decoder, where γ and β parameterise the affine transformation of style vectors.

Spatially-adaptive denormalisation. An alternative approach for combining content with style is the use of multiple Spatially-Adaptive Denormalisation (SPADE) [105] layers. As depicted in Fig. 3.4(c), a SPADE block receives the content channels and projects them onto an embedding space using two convolutional layers to produce the modulation parameters (tensors) γ and β. These parameters are then used to scale (γ) and shift (β) the normalised activations of the style representation.

3.4.3 Encouraging disentanglement in the latent space

The following operations and priors can be applied on a latent space to encourage disentanglement.

Gaussian prior. Encouraging the distribution of the encoded (vector) latent representation to match a Gaussian is a common prior. As reported in Section 3.3.1, such prior encourages the unsupervised disentanglement of the factors of variation and enables sampling for generating new images.

Task priors. As discussed in Section 3.3.3, content representation can be used for a downstream equivariant task e.g. semantic segmentation. Task losses, such as the segmentation loss, also contribute at learning a disentangled content representation [13]. Other task-based priors e.g. the number of human body parts [36], can be leveraged to encourage certain properties for the content.

Gradient reversal layer. The Gradient Reversal Layer (GRL) was introduced in [106] for domain adaptation, where the gradient is reversed to prevent the model from predicting undesired results. GRL is effective in learning domain-specific style representations [107]. Specifically, when using the style from one domain to generate images with style from another domains, the gradient is reversed to prevent this from happening.

Latent projection. Motivated by the findings of Michal et al. [81], which suggest that VAE encoders cannot model the arbitrary rotations of the representation space, Zhao et al. [108] propose the projection of the latent space onto the direction with more information about a generating factor. Latent projection allows the information to be disentangled between particular orientations of the data.

Frequency decomposition. Recent studies have investigated the use of frequency decomposition transformations to encourage CSD. For example, Liu et al. [109] use the fast Fourier transform to extract image amplitude and phase. Intuitively, the former reflects image style, whereas the latter corresponds to image content. Huang et al. [110] use Discrete Cosine Transformation (DCT) to extract the domain invariant and domain specific frequency components, as an approximation of content and style factors, respectively.

Structured latent. A causal approach to representation learning solves the identifiability problem by enforcing the latent space to be structured. Structured latents create strong inductive

biases because one might not only define the desired variables –which correspond to the generating factors– but also the relationship between them. This idea can be implemented in different settings, for example:

1. decomposing of a VAE latent space into separate parts, where each component is further processed at different levels of the decoder [111];

2. constraining the latent variable of a Bidirectional GAN (BiGAN) [112, 113] with Bayesian networks [114];

3. forcing the latent variables of a BiGAN-style architecture [115] to follow a graph structure prior defined as an adjacency matrix [116].

3.4.4 Learning setups for disentanglement

Popular learning setups can encourage disentanglement by harmonising the interaction between blocks.

Cycle-consistency. Cycle-consistency [117, 118, 119, 120] is a technique for regularising image translation settings. In particular, it can be useful for reinforcing correspondence between input and generated images [121, 122], or to improve stability and reconstruction fidelity in unsupervised and semi-supervised settings [123].

Latent regression. There is a gentle balance to be made in the complexity of these blocks: too complex and with lots of parameter capacity may lead to information captured within their parameters that can lead to this information not being captured in the latent variables. Latent regression has been employed to force the reconstructed image to contain information encoded into this representation [32]. In particular, considering an input image $\mathbf{X}$, the representation $\mathbf{z}$ and the reconstructed image $\mathbf{X}'$, we wish to extract a new latent representation $\mathbf{z}'$ from encoding $\mathbf{X}'$, which will be as similar as possible to $\mathbf{z}$. In other words, we need to minimise the distance between $\mathbf{z}$ and $\mathbf{z}'$ using a latent regression loss. The latent regression (LR) loss is defined as the ℓ_1 distance between $\mathbf{z}$ and $\mathbf{z}'$:

$$\mathcal{L}_{LR}(\mathbf{z}, \mathbf{z}') = |\mathbf{z} - \mathbf{z}'|_1, \tag{3.9}$$

where ℓ_1 distance is defined in Eq. 3.10.

3.5 Metrics for Disentanglement

To understand disentanglement and design models that improve it, we need to be able to quantify how disentangled is (are) the encoded representation(s). Below, I briefly report the most popular disentanglement metrics, splitting them into 2 categories: i) disentanglement of factors in a single vector latent variables, and ii) disentanglement between two latent variables of the same or different dimensionality.

Single vector-based latent variable. This category consists of both qualitative and quantitative methods for measuring how disentangled a representation is.

Qualitatively, we can evaluate disentanglement by traversing a single latent dimension that alters the reconstructed image by a single aspect (e.g. increase image intensity). In practice, these traversals are linear interpolations which are used to perform "walks" in non-linear data manifolds and to interpret the variation controlled by each factor [124, 125]. Latent traversals do not require ground truth information about the factors. Duan et al. [126] propose a way to quantify latent traversals in a post-hoc fashion, using the unsupervised disentanglement ranking metric to select the most disentangled version of the trained model. Quantitatively, there has been considerable effort to create metrics to evaluate vector representations. Since there are different proxies for disentanglement, popular metrics focus on measuring different aspects. For example, Higgins et al. [80] propose the first metric to quantify disentanglement when the ground truth factors of a data set are available. In fact, they evaluate disentanglement using the prediction accuracy of a linear classifier that is trained as follows: they first choose a factor k and generate data with this factor fixed, but all others varying randomly. After obtaining the representations of the generated data, they take the absolute value of the pairwise differences of these representations. Then, the mean of these statistics across the pairs gives one training input for the classifier, and the fixed factor index k is the corresponding training output. Subsequently, Kim and Mnih [127] adopt the metric of [80], but construct the training set of the linear classifier by considering the empirical variance of normalised representations rather than the pairwise differences. Chen et al. [128] argue that given a factor of variation, the first two dimensions of the latent vector should have the highest MI. They measure the gap between these two dimensions using the introduced mutual information gap metric. Ridgeway and Mozer [68] propose to measure the modularity of latent representations by measuring the MI between factors, ensuring that each vector dimension encodes at most one factor of variation. Eastwood and Williams [18] first train an encoder on a synthetic dataset with predefined

factors of variation z, and encode a representation c for each data sample. Then, they train a regressor to predict each factor z given a c representation. Based on the prediction accuracy, they measure the disentanglement, completeness, and informativeness of each representation. Finally, Kumar et al. [129] propose the separated attribute predictability score to first compute the prediction errors of the two most predictive latent dimensions for each factor, and then use the average error difference as a disentanglement metric. A more comprehensive review of metrics for vector-based disentanglement can be found in [130].

Two latent variables. The aforementioned metrics are not applicable in CSD as they rely on either having ground truth for the factors or assuming that the latent manifold is solely vector-based. To evaluate CSD one should consider more than one latent variable and a possible difference in dimensionality e.g. spatial content (tensor) and vector style. To the best of my knowledge, the only work that focuses on CSD metrics is that I will present in Chapter 4 published in [14]. In this work, I consider the properties of uncorrelation and informativeness, and propose to combine the empirical distance correlation [131] and a metric termed information over bias, to measure the degree of disentanglement *between* content and style representations. Two other methods for measuring the uncorrelation-independence between variables of different dimensionality are the kernel-target alignment [132] and the Hilbert-Schmidt independence criterion [133]. However, both methods require pre-defined kernels.

3.6 From Computer Vision to Medical Image Analysis

We are now well aware that learning disentangled representations requires supervision or design and learning biases. Using task-prior knowledge to incorporate proper biases to learn the desired disentangled representations is key for disentanglement in both domains. Medical applications can use, for instance, building blocks (Section 3.4) originally designed for computer vision tasks. One can also draw inspiration from how prior knowledge on the vision tasks has motivated the specific biases used. Below, with some exemplar computer vision tasks, I will discuss the connections between disentanglement in computer vision and medical domains.

3.6.1 Image-to-image translation

Image-to-image (I2I) translation aims to translate one image into another without changing the shape, i.e. content, which differs in a specific characteristic (e.g. style). A representative model

is MUNIT [102].

Connections to medical. Image-to-image translation in computer vision motivated many medical applications. In fact, several medical models are directly built based on MUNIT such as the ones in medical I2I translation [134], multi-modal and cross-modal segmentation [135], and registration [136]. The parallels here of domain-invariant spatial content and the domain-specific style representation, relate to separating anatomy and modality representations in the corresponding medical applications. A major difference though is that typically in medical image translation, we are particularly sensitive to maintaining identity when changing style. Several vision works show examples of day to night where content has changed slightly in the background. Such change will not be desired in medical tasks.

3.6.2 Facial attribute transfer

This task concerns the generation of a synthetic face that contains the target attribute, but without altering the subject identity (e.g. adding bangs to a subjects forehead). Most methods that focus on facial attribute transfer struggle with: a) transferring more than one attribute at a time, b) generating images based on exemplars, and c) achieving high-fidelity results. The first model to address the aforementioned challenges is ELEGANT [34], which encodes disentangled attribute representations of two exemplars in a vector latent space and performs attribute swapping. Apart from ELEGANT, Lin et al. [35] propose a GAN model with a domain classifier to learn to transfer attributes between multiple domains. He et al. [137] present a GAN that conditions the face generation of opposite samples (e.g. smile, no smile) using one-hot attribute vectors. Zhou et al. [138] exploit cycle consistency to transfer attributes, with the limitation that the attributes should have approximately the same spatial location.

Connections to medical. When transferring facial attributes, the subject identity should be preserved and only some attributes transferred. This transferral is desirable in several medical applications such as brain aging [121] and controllable synthesis [139], where the synthesised brain or heart images should contain the identity information of the original images but with different ages or pathology. ELEGANT preserves the identity information by only modifying the local part of the image. The medical models similarly modify the local anatomy parts but also apply the identity or consistency losses to the remaining parts of the image. We should note that most face models rely on pre-trained or pre-extracted strong priors to identify facial features. Such strong priors are rarely available in medical imaging.

3.6.3 Pose estimation

For pose estimation, the human body constitutes a strong content prior that can be exploited to encode body structure in a spatial and semantic latent space, to be used for equivariant tasks that require body joint position. Lorenz et al. [36] propose to apply the equivariance and invariance losses to learn the equivariant (content) and invariant (style) representations and use this type of disentanglement for this challenging articulated body pose estimation task. Esser et al. [37] adopt the disentanglement of the human body pose from the corresponding appearance (style) information in the context of a dual-encoder VAE setting, where they use the body-related factors for human appearance transfer and synthesis [140].

Connections to medical. Similar to the human body, human organs *e.g.* brain and heart, have strong anatomical structure priors, which can be similarly used for learning disentangled representations with equivariance and invariance properties. For example, similar to the invariance loss in [36], Bercea et al. [141] apply the shape consistency loss to encourage the shape embeddings of brain MRI images to be invariant to Gamma shifts. However, it is not always possible to assume such strong structural priors as diseases or abnormalities exist.

3.7 Compositionality

Lastly, I review the compositionality in modern machine learning and deep learning. Compositionality is a fundamental concept in computer vision, where it refers to the ability to recognize complex objects or scenes by combining simpler components or features [142]. Recently, there has been a growing interest in developing models that can effectively capture the compositional nature of visual information, leading to improved performance on various vision tasks [143].

To integrate compositionality, compositional representation learning is an area of active research in computer vision [144] and natural language processing [145]. Intuitively, learning a good latent representation should also reflect the compositional property if the compositional structure is exhibited in the input data [146]. The compositionality of latent space has already been explored extensively in language processing because of the compositionality of languages [147]. Recently, the language processing community has started to explore whether compositionality arises in learning problems where the compositional structure has not been built in from the start [148].

In addition, there is also some research about using compositionality in computer vision [149, 150]. Compositional representation learning in computer vision involves learning representations of complex objects by combining representations of their constituent parts [143, 151, 152]. The goal is to learn representations that are robust to changes in the appearance of the object and that can generalise to new objects [153, 154, 155]. Early approaches to compositional representation learning in computer vision include the bag-of-visual-words model [156] and part-based models [149]. The bag-of-visual-words model involves representing an image as a histogram of visual words, which are learned from a training set of images. Part-based models involve learning representations of objects by decomposing them into parts and learning representations of the parts and their relationships. Compositional representation learning has been applied to fine-grained recognition tasks in computer vision, such as recognizing bird species [152, 157]. These tasks require learning representations that capture subtle differences between objects and their parts. Compositional representation learning has been shown to be effective for fine-grained recognition by allowing the model to reason about the relationships between parts and the whole object. In addition, compositionality has been also incorporated for robust image classification [153, 149] and recently for compositional image synthesis [158, 159]. Among these work, Compositional Networks [149] originally designed for robust classification under object occlusion is easier to extend to pixel-wise tasks as it learns spatial and interpretable vMF likelihoods. Previous work integrates the vMF kernels and likelihoods [149] for object localisation [160] and recently for nuclei segmentation (with the bounding box as supervision) in a weakly supervised manner [161].

3.8 Training Losses

In the thesis, several training losses are frequently used. Here, I define these losses with the example of considering the inputs for the loss functions as 2D images or segmentation masks i.e. $\mathbf{X}$ the ground truth and $\hat{\mathbf{X}}$ the prediction. H and W are the height and width of $\mathbf{X}$.

The ℓ_1 distance is defined as:

$$\mathcal{L}_{\ell_1}(\mathbf{X}, \hat{\mathbf{X}}) = |\mathbf{X} - \hat{\mathbf{X}}|_1 = \frac{1}{H * W} \sum_{h \in H} \sum_{w \in W} |\mathbf{X}(h, w) - \hat{\mathbf{X}}(h, w)|. \tag{3.10}$$

The Mean Squared Error (MSE) is defined as:

$$\mathcal{L}_{MSE}(\mathbf{X}, \hat{\mathbf{X}}) = |\mathbf{X} - \hat{\mathbf{X}}|_2 = \frac{1}{H * W} \sum_{h \in H} \sum_{w \in W} (\mathbf{X}(h, w) - \hat{\mathbf{X}}(h, w))^2. \tag{3.11}$$

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

$$\mathcal{L}_{Dice}(\mathbf{X}, \hat{\mathbf{X}}) = 2\frac{|\mathbf{X} \cap \hat{\mathbf{X}}|}{|\mathbf{X}| + |\hat{\mathbf{X}}|}. \tag{3.12}$$

The Kullback–Leibler divergence (KLD) loss is defined as:

$$\mathcal{L}_{KL}(p(\mathbf{X})||p(\mathbf{z})) = \mathbb{E}_{p(\mathbf{X})}[\log p(\mathbf{X}) - \log p(\mathbf{z})]. \tag{3.13}$$

3.9 Summary

In this chapter, I introduced the key concepts of representation learning, focusing on the definition of disentanglement, generative models for learning disentangled representations, the fundamental disentanglement building blocks, and metrics of disentanglement. I also discussed what we can learn from computer vision for medical image analysis tasks. Then, I introduced compositionality which is detailed in Chapter 6. Finally, several training losses used in the thesis are defined. In the next chapter, I will discuss how different learning and design biases including those discussed in this chapter affect disentanglement.

Chapter 4
Metrics for Exposing the Biases of Content-Style Disentanglement

4.1 Introduction

In this chapter, I will describe how extensive inductive biases and learning biases (mostly revised in Chapter 3) are applied to achieve certain disentanglement. To fully understand how disentangled representations are learnt, I conduct a comprehensive review of popular disentanglement models in different applications and summarised the frequently used building blocks. I further study content-style models by using two proposed metrics to measure the degree of disentanglement between content and style representations.

4.1.1 Motivation of the approach

Recent work in representation learning argues that to achieve explainable and compact representations, one should separate out, or disentangle, the underlying explanatory factors into different dimensions of the considered latent space [28, 57]. In other words, it is beneficial to obtain representations that can separate latent variables that capture sensitive and useful information for the task at hand from the less informative ones [56]. Disentanglement has recently been shown to improve task performance, model generalisation, and representation interpretability [162, 163, 164, 165, 166, 79, 167, 140]. Unfortunately, disentangling without supervision is an ill-posed and impossible task [31, 168, 169] and, to obtain it, we must introduce restrictions and inductive priors [31, 168]. These priors are different forms of "bias" imposed by model design (design bias), learning objectives (learning bias), and data (data bias).

This chapter is based on:

- **Liu, X.***, Thermos, S.*, Valvano, G.*, Chartsias, A., O'Neil, A. and Tsaftaris, S.A., 2021. Measuring the Biases and Effectiveness of Content-Style Disentanglement. British Machine Vision Conference 2021. *Equal contribution.

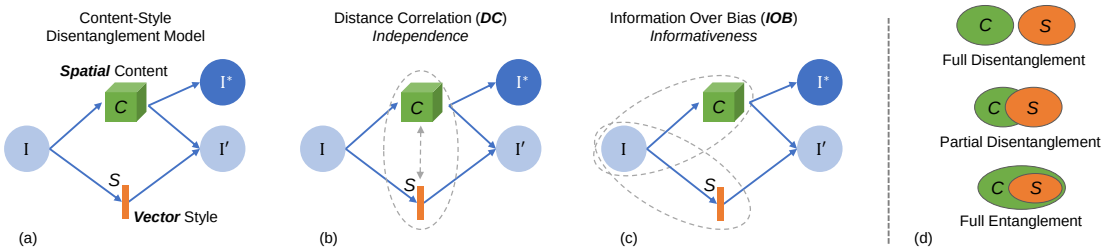


Figure 4.1: **(a)** A schematic representation of disentanglement between spatial content C and vector style S in the context of a primary and a secondary spatially equivariant task ($\mathbf{I}'$, $\mathbf{I}^*$). Measuring the degree of C-S disentanglement using distance correlation **(b)** and information encoded over the input bias **(c)**. **(d)** A visual description of degrees of C-S (dis)entanglement. This figure was originally produced by Dr. Spyridon Thermos. Reproduced with permission.

In this chapter, I set out to reveal such choices of bias in state-of-the-art (SOTA) disentanglement methods. The particular focus is on "content-style" disentanglement, which decomposes input images into spatial "content" and vector "style" representations. In principle, content (C) should contain the semantic information required for spatially equivariant tasks (e.g. segmentation and pose estimation), whereas style (S) contains information on image appearance (e.g. color intensity and texture). However, contrary to extensive research on quantifying the degree of disentanglement between vectors [129, 128, 18, 68, 84, 170, 171], there is no analysis of C-S disentanglement. In fact, to the best of my knowledge, there is no study identifying the training biases enforced in C-S disentanglement settings or exposing the true relationship between the degree of disentanglement and model performance.

4.1.2 Approach overview

The overview of the approach is illustrated in Fig. 4.1. Considering the C-S disentanglement framework, I and the co-authors propose two complementary metrics to evaluate two properties in the context of C-S disentanglement: *(un)correlation*, and *informativeness*.

4.1.3 Contributions

Herein, I attempt to bridge these gaps with the **contributions**:

- I identify and analyse the key biases in SOTA models that employ C-S disentanglement.

I show how the biases affect disentanglement and task performance (utility) in three popular vision tasks: image translation, segmentation, and pose estimation.

- To make a quantitative analysis possible, I propose two complementary metrics building on existing work, to evaluate C-S disentanglement in terms of the amount of information encoded in each latent variable (informativeness) and (un)correlation between the encoded *spatial tensor* content and *vector* style (a proxy for independence).
- I find that: a) lower C-S disentanglement benefits task performance if a specific style-related prior is not violated; and b) performance is highly correlated with latent variable informativeness. I also assess content semanticness (interpretability).

This chapter is organised as follows. Section 4.2 mentions previous work related to ours. Section 4.3 describes the proposed metrics. Section 4.4 presents the validation of the effectiveness of the proposed metrics with the toy dataset. Section 4.5 talks about the considered applications and models. Section 4.6 shows the experiments and results. Section 4.7 discusses the correlation of the metrics. Section 4.8 answers the key questions and presents the major findings. Finally, this chapter is concluded in Section 4.9.

4.2 Related Work

Content-Style disentanglement. Image-to-Image translation has extensively explored the decoupling of image style and content [172, 33, 32, 173]. Content-style disentanglement was also used in other applications, such as semantic segmentation [13] and pose estimation [174], where the content serves as a robust representation for downstream tasks. In general, most methods derive latent spaces capturing C or S information using auto-encoder variants.

These models achieve C-S disentanglement through different biases, such as architectural choices (e.g. AdaIN [102], content binarization [13]), learning objectives (e.g. Kullback-Leibler divergence, latent regression loss, de-correlation losses in vector representations [175, 176]), or supervisory signals (e.g. using content for segmentation [13]). However, the precise effect of each bias on disentanglement and model performance is not thoroughly explored.

Evaluating disentanglement. Recently, several methods have been proposed for assessing the degree of disentanglement in a vector latent variable. A classical approach is *latent traversals*: a visualization showing how traversing single latent dimensions generates variations in the

image reconstruction. Latent traversals do not need ground truth information on the factors, and can be used in mixed tensor spaces [13, 36] to offer qualitative evaluations. Alternatively, latent traversals can be combined with pre-trained networks to measure the perceptual distance between the produced embeddings [84].

There exist several ways in quantitatively evaluating representations learned by VAEs and GANs. Unfortunately, these methods rely only on vector representations, and some also peruse ground truth knowledge about the latent factors. In particular, some methods try to associate known factors of variations (e.g. rotation) with specific latent dimensions [80, 127] or manifold topology [177]. Others measure the ability to isolate one factor in a single vector latent variable [129], measuring compactness or modularity [128, 18, 170], linear separability [84], consistency and restrictiveness [178], and explicitness of the representation [68]. Lastly, there is work on measuring the factor informativeness in a vector latent variable w.r.t. the input, independence among factors, as well as interpretability [171, 18].

The aforementioned metrics cannot be directly employed to C-S disentanglement settings, where the latent factors have different dimensionality (e.g. the style is a vector and the content a spatial multi-channel tensor). However, in this chapter I attempt to transfer these concepts to the C-S disentanglement domain, incorporating both spatial (tensor) and vector representations[1] to expand our understanding of the relation between C-S disentanglement and: a) biases adopted by each model; b) task performance; c) representation interpretability.

4.3 Measuring Properties of Disentangled Content and Style

Given N image samples $\{\mathbf{I}_i\}_{i=1}^N$, I assume two representations of content and style: $\{\mathbf{C}_i\}_{i=1}^N$ and $\{\mathbf{s}_i\}_{i=1}^N$, respectively. Building on existing work in vector-based disentanglement [18, 171], I present two complementary metrics to evaluate two properties in the context of C-S disentanglement: *(un)correlation*, and *informativeness*. I provide evidence that the metrics offer complementary information in Section 4.7. Then, I discuss two properties of the disentangled representations, namely their *utility* and *interpretability*.

Distance Correlation (DC). Disentangled representations separate content and style into independent latent spaces [57], satisfying $p(\mathbf{C}, \mathbf{s}) = p(\mathbf{C})p(\mathbf{s})$.However, directly measuring in-

[1]Note that the metrics used for the analysis are generic and can be readily applied to vector-based C-S disentanglement methods, such as [89].

dependence between spatial C and vector S with existing metrics is not feasible. Since independent representations (variables) must be uncorrelated [128, 179], I use the *empirical Distance Correlation* (DC) [131] to measure the correlation of distance between tensors of arbitrary dimensionality. Note that DC is bounded in the $[0, 1]$ range, while differently from other correlation-independence metrics, such as the kernel target alignment [132] and the Hilbert-Schmidt independence criterion [133], it has the advantage of not requiring any pre-defined kernels. Moreover, DC is 0 if and only if the random variables are independent. Thus, distance correlation measures both linear and nonlinear relationships between two random variables.

For N samples, consider two N-row matrices $\mathbf{T}_1$ and $\mathbf{T}_2$. In general, $\mathbf{T}_1$ and $\mathbf{T}_2$ row dimension varies as they are formed by concatenating images $\mathbf{I}_i$, content features $\mathbf{C}_i$ or style features $\mathbf{s}_i$. For $\mathbf{I}_i$ and $\mathbf{C}_i$ I first concatenate the channels and then row-scan to form a vector; $\mathbf{s}_i$ is already a vector. DC is then defined as:

$$DC(\mathbf{T}_1, \mathbf{T}_2) = \frac{dCov(\mathbf{T}_1, \mathbf{T}_2)}{\sqrt{dCov(\mathbf{T}_1, \mathbf{T}_1) dCov(\mathbf{T}_2, \mathbf{T}_2)}}, \text{ with } dCov(\mathbf{X}, \mathbf{Y}) = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{\mathbf{A}_{i,j}\mathbf{B}_{i,j}}{N^2}}. \tag{4.1}$$

Here, $dCov$ is the distance covariance between any two N-row matrices $\mathbf{X}$ and $\mathbf{Y}$, while $\mathbf{A}$ and $\mathbf{B}$ are their respective distance matrices. In particular, each matrix element $\mathbf{A}_{i,j}$ of $\mathbf{A}$ is the Euclidean distance between two samples $\|\mathbf{X}^i - \mathbf{X}^j\|$, after subtracting the mean of row i and column j, as well as the matrix mean. Formally, we define the distance matrix as:

$$\begin{aligned} a_{i,j} &= \|\mathbf{X}^i - \mathbf{X}^j\|, \quad \bar{a}_{i,\cdot} = \frac{1}{N}\sum_{i=1}^{N} a_{i,j}, \quad \bar{a}_{\cdot,j} = \frac{1}{N}\sum_{j=1}^{N} a_{i,j}, \\ \bar{a}_{\cdot,\cdot} &= \frac{1}{N^2}\sum_{i,j=1}^{N} a_{i,j}, \quad \mathbf{A}_{i,j} = a_{i,j} - \bar{a}_{i,\cdot} - \bar{a}_{\cdot,j} - \bar{a}_{\cdot,\cdot}. \end{aligned} \tag{4.2}$$

$\mathbf{B}$ is similarly calculated for $\mathbf{Y}$. I estimate disentanglement between C and S using distance correlation, $DC(\mathbf{C}, \mathbf{s})$, with values closer to 0 indicating higher disentanglement. C and S can be uncorrelated, e.g. $DC(\mathbf{C}, \mathbf{s}) = 0$, either when they encode unrelated information or when one encodes *all* information and the other encodes *noise*. The latter indicates posterior collapse, thus full entanglement. To tackle this, $DC(\mathbf{C}, \mathbf{s})$ needs a complementary metric to measure the representations' informativeness.

Information Over Bias (IOB). To measure the amount of information encoded in C and S, I

introduce the *Information Over Bias* (IOB) metric, aiming to detect posterior collapse when C and S are disentangled, but one (C or S) is not informative about the input. Given $\mathbf{z} \in \{\mathbf{C}, \mathbf{s}\}$ produced from N images at inference, I measure the amount of information encoded in each representation. I train a decoder $\boldsymbol{G}_{\theta_n}$, a neural network with parameters θ_n, to reconstruct images $\mathbf{I}$ by minimising the Mean Squared Error (MSE) between the reconstructed images and the original images, given $\mathbf{z}$. Each decoder is trained for 40 epochs with batch size 10 using Adam optimiser [180]. Using MSE to measure the quality of the reconstruction, we can evaluate how informative $\mathbf{z}$ is with respect to the image $\mathbf{I}$. In addition, any network design for the decoder will introduce biases. To de-bias, we also train a decoder to reconstruct images with a fixed input, where all the elements are 1. IOB is defined as the expectation over the test images of the ratio:

$$IOB(\mathbf{I}, \mathbf{z}) = \mathop{\mathbb{E}}_{i}\left[\frac{\mathrm{MSE}(\mathbf{I}_i, \boldsymbol{G}_{\theta_1}(\mathbb{1}))}{\mathrm{MSE}(\mathbf{I}_i, \boldsymbol{G}_{\theta_2}(\mathbf{z}_i))}\right] = \frac{1}{N}\sum\nolimits_{i=1}^{N}\left(\frac{\frac{1}{K}\sum_{k=1}^{K}||\mathbf{I}_i^k - \boldsymbol{G}_{\theta_1}(\mathbb{1})||^2}{\frac{1}{K}\sum_{k=1}^{K}||\mathbf{I}_i^k - \tilde{\mathbf{I}}_i^{\,k}||^2 + \varepsilon}\right), \quad (4.3)$$

where $\mathbf{I}$ and $\tilde{\mathbf{I}}$ are an image and its reconstruction obtained through $\boldsymbol{G}_{\theta_n}$; $i = 1 \dots N$, $k = 1 \dots K$, $n = 1 \cdots + \infty$ are indices iterating on the test images, the image pixels, and the generator model index (different for each run); ε is a small value that prevents division by zero. Note that the ratio aims at ruling out from IOB both data correlations (common structure, colors, pose, etc., across the images of the dataset) and architectural biases that one could introduce in the design of $\boldsymbol{G}_{\theta_n}$. In particular, this is done by computing the ratio between the MSE obtained after training $\boldsymbol{G}_{\theta_n}$ to reconstruct the images from their *informative* representation $\mathbf{z}$ (e.g. $\mathrm{MSE}(\mathbf{I}_i, \boldsymbol{G}_{\theta_2}(\mathbf{z}_i))$, and after training $\boldsymbol{G}_{\theta_n}$ from an *uninformative* constant tensor $\mathbb{1}$ (e.g. $\mathrm{MSE}(\mathbf{I}_i, \boldsymbol{G}_{\theta_1}(\mathbb{1}))$). In the latter case, $\boldsymbol{G}_{\theta_n}$ will only learn the dataset bias it can model, given θ_n. Hence, high values of IOB can be associated with higher information inside the representation $\mathbf{z}$, while the lower bound $IOB = 1$ means that no information of the images $\mathbf{I}$ is encoded in $\mathbf{z}$.[2]

Utility and interpretability. As discussed, I can use DC and IOB to measure the degree of disentanglement between latent representations. However, one of the primary goals of disen-

[2]Optimising $\boldsymbol{G}_\theta$ with stochastic gradient descent can introduce noise and slightly alter the measure. For example, IOB may, in practice, even be slightly smaller than 1. Thus, I average results across multiple runs and initializations of $\boldsymbol{G}_\theta$, which contributes to the computational load of estimating IOB.

tanglement is to improve task performance (utility) and representation interpretability, hence I also investigate the relationship between C-S disentanglement and these two notions. In particular, I measure utility by quantifying performance on a downstream task, which for disentangled representations is typically image translation [32, 33] to translate image content from one domain to another. I also consider tasks using content e.g. to extract segmentations [13] or landmarks [36], and therefore assess how effectively it can be used in downstream tasks. I detail performance metrics for each application in Section 4.6.

Assessing interpretability is not trivial. Here, I assume that interpretability implies semantic representations. Previously, vector representations were considered semantic if a portion of the latent space corresponded to specific data variations [181, 182]. Style semantics were qualitatively evaluated with latent traversals of individual dimensions [13]. Thus, I consider a style interpretable if images produced by linear traversals in the style latent space are realistic and smoothly change intensity. In spatial representations, such data variation should be confined to individual objects: thus, semantic content should split distinct objects into separate channels of C. Wherever possible, I evaluate this with qualitative visuals.

4.4 Validating the Effectiveness of DC and IOB

To verify the effectiveness of DC and IOB, I design an experiment using the synthetic teapot dataset [18], which consists of 200k of 64×64 pixel resolution images of a teapot with varying pose and colour. Each image of this dataset is generated using 5 ground truth (GT) generating factors (scalars), e.g. azimuth, elevation, red, green, and blue colour, independently sampled from 5 different uniform distributions. I consider the 3 color factors as the GT style (GT S) representation, while as GT spatial content representation (GT C) I leverage the segmentation mask of the object, which correlates with the azimuth and elevation factors (for visual examples see Fig 4.2).

As the content and style representations for the teapot dataset are independent, I first evaluate DC and IOB using the GT C and S representations, and the input images, which are expected to reflect the independence between the GT C and S representations, hence justifying the desired properties of the metrics i.e. measuring (un)correlation. Then, I sample from a uniform distribution $U[0, 1]$ to generate a new, random style and content representations for each image, and evaluate the metrics using the following scenarios: a) random content, GT style and

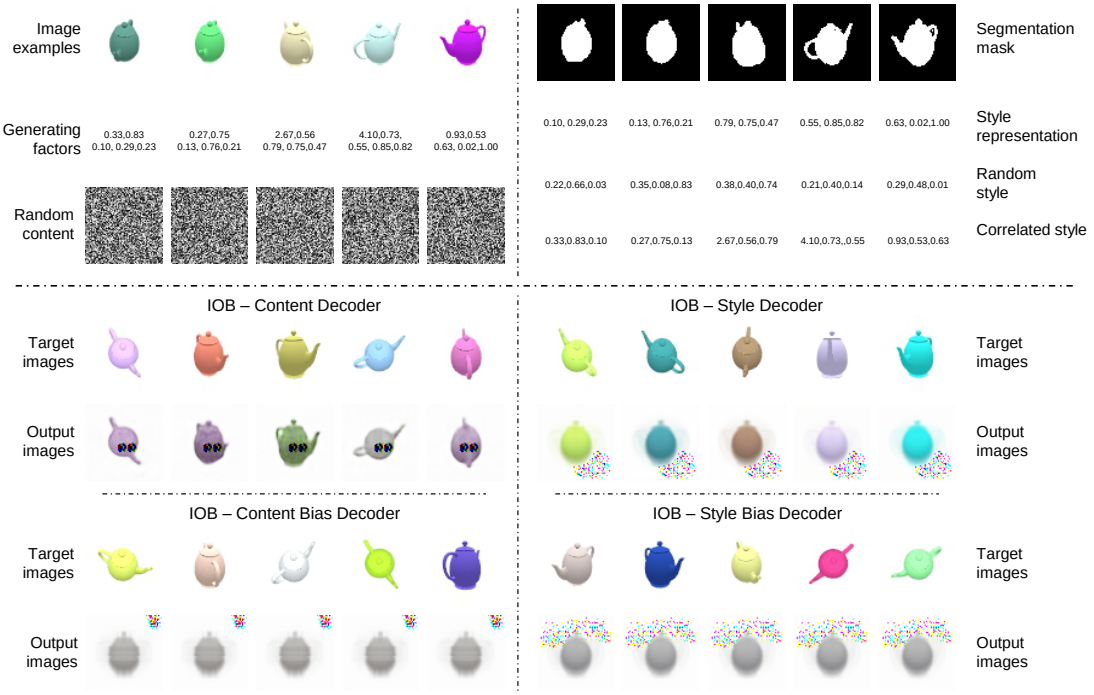


Figure 4.2: Visuals for the empirical study with the teapot dataset. Top: examples of original images, ground truth generating factors and segmentation masks. I also show the randomly sampled content and style representations. Bottom: examples of target images and output images for the IOB decoders. The artefacts in the reconstructed images indicate the biases introduced by the network design. Figure is taken from [14].

images; b) GT content, random style and images; c) random content, random style and images. These scenarios simulate that the models learn disentangled representations that are not informative to the image i.e. posterior collapse cases. Finally, to approximate the highly entangled C and S scenario, I construct the content-correlated style representations (correlated S) as the azimuth, elevation and red colour factors. For each experiment, I randomly sample 5k images and the GT representations, while all results are the average of 3 different runs.

For the structure of IOB decoders, I vary the number of layers for different applications due to the different dimensions of the representations. The design of the decoders for the teapot dataset can be found in Table 4.1. Overall, $\boldsymbol{G}_\theta(\mathbf{s})$ consists of several linear layers, followed by transpose (upsampling steps) and one plain CONV layer that generates the final image. $\boldsymbol{G}_\theta(\mathbf{C})$ follows an autoencoder structure with several encoder and decoder CONV layers. For the teapot dataset, the content representation has a size of $1 \times 64 \times 64$ and the style representation has a

Decoder	Input Shape→Output Shape	Layer Information
$\boldsymbol{G}_\theta(\mathbf{C})$	(1,64,64)→(8,64,64)	CONV-(O:8,K:7x7,S:1,P:3), IN, Leaky ReLU
	(8,64,64)→(16,32,32)	CONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,32,32)→(32,16,16)	CONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,16,16)→(64,8,8)	CONV-(O:64,K:4x4,S:2,P:1), IN, Leaky ReLU
	(64,8,8)→(32,16,16)	DECONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,16,16)→(16,32,32)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,32,32)→(8,64,64)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,64,64)→(3,64,64)	CONV-(O:3,K:7x7,S:1,P:3), Tanh
$\boldsymbol{G}_\theta(\mathbf{s})$	(3)→(256)	FC-(O:256)
	(256)→(4096)	FC-(O:4096), Flatten
	(64,8,8)→(32,16,16)	DECONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,16,16)→(16,32,32)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,32,32)→(8,64,64)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,64,64)→(3,64,64)	CONV-(O:3,K:7x7,S:1,P:3), Tanh

Table 4.1: IOB decoders design for the teapot dataset. The notations in the tables are: O: the number of output channels; K: the kernel size; S: the stride size; P: the padding size; FC: fully-connected layer; IN: instance normalisation;

size of 3.

From Table 4.2, I observe that for any combination of C and S (except for the correlated S one), the $DC(\mathbf{C}, \mathbf{s})$ is low, which indicates that the representations are highly uncorrelated. This result meets the expectation as the colour (S) and the azimuth or the elevation factors (C) are independent in the teapot dataset. However, I also observe a high $DC(\mathbf{C}, \mathbf{s})$ value, e.g. 0.53, between GT C and correlated S, which verifies that DC can indeed detect the entangled representations case. Additionally, the effectiveness of the DC metric is validated by the high $DC(\mathbf{I}, \mathbf{C})$ values when using GT C representations, versus the low values when using random C ones. Note that the DC between the GT S and image is higher than the one between GT C and image, which is reasonable as S and image have nearly one-to-one mapping relationship, while the segmentation masks for different images can be similar. The IOB results, reported in Table 4.2, also reflect that the segmentation mask is less informative ($IOB(\mathbf{I}, \mathbf{C}) = 1.73$) about the input image compared to S ($IOB(\mathbf{I}, \mathbf{s}) = 2.47$) for the GT C and GT S case. This is a result of the strong dataset bias, where given that the object is always a teapot, it is the colour of the reconstructed image that makes it more similar to the input one in terms of MSE.

Metric	GT C GT $\mathbf{s}$	Random C GT $\mathbf{s}$	GT C Random $\mathbf{s}$	Random C Random $\mathbf{s}$	GT C Correlated $\mathbf{s}$
$DC(\mathbf{C}, \mathbf{s})$ ($\downarrow$)	0.17 $\pm$0.00	0.13 $\pm$0.04	0.05 $\pm$0.00	0.13 $\pm$0.04	0.53$\pm$0.02
$DC(\mathbf{I}, \mathbf{C})$ ($\uparrow$)	0.64 $\pm$0.03	0.16 $\pm$0.05	0.64 $\pm$0.03	0.16 $\pm$0.05	0.64$\pm$0.03
$DC(\mathbf{I}, \mathbf{s})$ ($\uparrow$)	0.87 $\pm$0.00	0.87 $\pm$0.00	0.04 $\pm$0.00	0.04 $\pm$0.00	0.33$\pm$0.00
$IOB(\mathbf{I}, \mathbf{C})$ ($\uparrow$)	1.73 $\pm$0.10	1.41 $\pm$0.20	1.73 $\pm$0.10	1.41 $\pm$0.20	1.73$\pm$0.10
$IOB(\mathbf{I}, \mathbf{s})$ ($\uparrow$)	2.47 $\pm$0.78	2.47 $\pm$0.78	0.76 $\pm$0.15	0.76 $\pm$0.15	2.70$\pm$0.26

Table 4.2: Empirical study results for the DC and IOB metrics evaluation using the teapot dataset [18]. Results are in "mean $\pm$std" format.

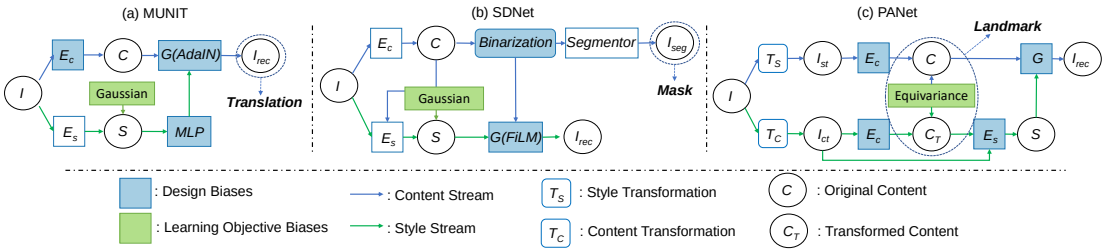


Figure 4.3: Model schematics. a) MUNIT: Instance normalisation is used to remove style from content; $\boldsymbol{E}_s$ uses global pooling. b) SDNet: the content is represented with binary features; style is forced to approximate a normal prior. c) PANet: content and style are encouraged to be equivariant to intensity and spatial transformations. Figure is taken from [14].

Visual examples and qualitative results of the empirical study on the proposed metrics with the teapot dataset are included in Fig. 4.2. It is notable that the artifacts in the reconstructed images introduced by the decoder bias are observed in the results of both decoders and bias decoders, which demonstrates the motivation of de-biasing design in IOB.

4.5 Considered Vision and Medical Applications

Many applications disentangle C from S [183, 89, 184, 90] or other attributes, such as pose, geometry, and motion [185, 186, 187, 188], to improve performance in vision tasks. For the analysis, I select and discuss three popular approaches (see Fig. 4.3) from diverse applications, namely image translation (MUNIT [32]), semantic segmentation (SDNet [13]), and pose estimation (PANet [36]). All resemble auto-encoders, mapping input images to disentangled

features but use several biases, which are detailed below. The scope is to elucidate how each bias affects disentanglement using these models and their chosen biases as exemplars.

4.5.1 MUNIT for image-to-image translation

Multimodal Unsupervised Image-to-image Translation (MUNIT) [32] does not impose strict constraints on the learned representations, and achieves disentanglement with both design and learning biases. The model is depicted in Fig. 4.3(a).

The basic assumption is that multi-domain images (a necessary *data bias*), share common content information, but differ in style. A content encoder maps images to multi-channel feature maps, by removing style with Instance Normalisation (IN) layers [102] (*design bias*). A second encoder extracts global style information with fully connected layers and global pooling. Finally, style and content are combined in a decoder with AdaIN modules [102] (*design bias*).

Disentanglement is additionally promoted with a bidirectional reconstruction loss [189] that enables style transfer. To learn a smooth representation manifold, two Latent Regression (LR) losses (*learning bias*) are applied: content LR penalizes the distance to the content extracted from reconstructed images, whereas style LR encourages encoded style distributions to match their Gaussian priors. Finally, adversarial learning encourages realistic synthetic images.

4.5.2 SDNet for medical image segmentation

SDNet [13] is a semi-supervised framework that disentangles medical images in anatomical features (content) and imaging-specific characteristics (style). The model is depicted in Fig. 4.3(b). Similarly to other models, SDNet uses separate content and style encoders, but here a segmentation network is applied on the content features trained with supervised objectives and annotated images (*data bias*).

However, in contrast to MUNIT, SDNet does not impose a design bias on the encoder, but rather on the content which is represented as multi-channel binary maps of the same resolution as the input (*design bias*).

This is obtained with a softmax and a thresholding function with the straight-through operator [190], such that any style is removed from the content. To encourage style features to encode residual information (and not content), a loss enforces the style representation to ap-

proximate a standard Gaussian, following the VAE formulation [30] (*learning bias*). In this setup, any information encoded in style comes at a cost, and thus encoding redundant information is prevented [191]. Furthermore, a LR loss of the style is employed to prevent posterior collapse of the decoder (*learning bias*).

Finally, style and content are combined to reconstruct the input image by applying a series of convolutional layers with feature-wise linear modulation (FiLM) conditioning. Similarly to AdaIN, FiLM modules are restrictive, allowing the style only to normalise the conditioned feature maps, and thus further discouraging the style from encoding content information (*design bias*).

4.5.3 PANet for pose estimation

For the pose estimation task, I consider a dual-stream autoencoder denoted as Pose Appearance Network (PANet) [36]. PANet consists of two branches that decouple pose (content) and appearance (style) but employs heavily entangled encoders-decoders. The model is depicted in Fig. 4.3(c).

The content is represented as a multi-channel feature map, where each channel corresponds to a specific body part (since the number of parts are fixed, this imposes a strong *data bias*). A Gaussian distribution is applied to each feature map to remove any style information, whilst also preserving the spatial correspondence (*design bias*).

The corresponding style information is extracted from the encoder features using average pooling (*design bias*). More critically, style vectors do not correspond to global image style, since they are applied to specific content parts during decoding (*design bias*).

Finally, disentanglement is encouraged with a transformation equivariance loss (*learning bias*). This ensures that the spatial transformations, such as translations and rotations, affect only the content, while the intensity ones, such as the color and texture information, affect only the style.

4.5.4 Summary of the applications

Table 4.3 summarizes the *design* and *learning biases* of the methods. Note that the biases are reported as modules, without indicating the way they are used in the experiments (e.g. AdaIN is reported without specifying that it is removed from the original MUNIT, but is added to PANet

		MUNIT	**SDNet**	**PANet**
Design Bias	AdaIN	√		√
	Instance Normalisation	√		
	SPADE		√	
	Binarization		√	
	MLP			√
Learning Bias	Latent Regression	√	√	
	KL Divergence		√	
	Equivariance			√

Table 4.3: Overview of the *design* and *learning biases* that are investigated in the context of the three investigated vision tasks: a) image-to-image translation (MUNIT), b) medical segmentation (SDNet), and c) pose estimation (PANet). Note that the biases here specifically mean model designs or learning objectives.

as a variant).

4.6 Experimenting on Vision and Medical Applications

Here I briefly summarise how each bias is enforced, whilst the detailed model descriptions and a summary of their *design* and *learning* biases can be found in Section 4.5. In particular, for: **a) MUNIT** I consider ablations removing IN [192], AdaIN layers, or style LR loss (for fairness, I do not remove LR of the content as it is fundamental for the functioning of the model); **b) SDNet** I identify content binarization, Gaussian approximation, LR and the FiLM-based [96] decoder as the main biases that affect C-S disentanglement. I investigate their impact on the representations and their effect on semantic segmentation; **c) PANet** I remove the Gaussian prior and replace its specific C-S conditioning with AdaIN. I analyse PANet performance in pose estimation. These models help us cover the following diverse cases: **i)** no supervision and weak C constraints (MUNIT), **ii)** no supervision with strong C constraints (PANet), and **iii)** supervision with strong C constraints (SDNet).

General setup. For each model, I analyse the effect that design choices and learning objectives have on disentanglement and task performance, and I evaluate utility and interpretability of the

learned representations. I use the implementations provided by the authors, ablating only the components needed for the analysis. In all tables, arrows $(\uparrow, \downarrow)$ indicate direction of metric improvement; best results are in bold. Numbers are the average of 5 different runs.

4.6.1 Model design and training scheme for IOB

The design of the decoders can be found in Table 4.4, Table 4.5, Table 4.6. Overall, $\boldsymbol{G}_\theta(\mathbf{s})$ consists of several linear layers, followed by transpose (upsampling steps) and one plain CONV layer that generates the final image. $\boldsymbol{G}_\theta(\mathbf{C})$ follows an autoencoder structure with several encoder and decoder CONV layers. For MUNIT, the content representation has size $128 \times 64 \times 64$ and the style representation has size 8. For SDNet, the content representation has size $8 \times 224 \times 224$ and the style representation has size 8. For PANet, the content representation has size $3 \times 64 \times 64$ and the style representation has size 1024. Note that it is not necessary to have exactly same design as in the tables, where the key suggestion is to design the decoders to generate as high-quality as possible reconstructed images.

All the decoders are trained using the Adam optimiser [180] ($\beta_1 = 0.5, \beta_2 = 0.999$) with a learning rate of $1e^{-4}$ for 40 epochs using batch size 10.

4.6.2 Image-to-image translation

I consider the original MUNIT and three variants: **i)** I replace the AdaIN modules of the decoder with simple style concatenations, reducing the restrictions on the re-combination of C and S. **ii)** I remove the LR loss, responsible for the style following a Gaussian. **iii)** I remove IN from the content encoder, to confirm that it helps to cancel out original style and retain the content only [102]. As [32] I evaluate quality and diversity of the translated images using the Fréchet Inception Distance (FID) [193] and (Learned Perceptual Image Patch Similarity) LPIPS [194].

Data. I use SYNTHIA [19], which consists of over $20,000$ rendered images and corresponding pixel-level semantic annotations, where 13 classes of objects are labeled for aiding segmentation and scene understanding problems. I also use Cityscapes [20], which contains a set of diverse street scene stereo video sequences and over 5k frames of high-quality semantic annotations, where 30 classes of instances are labeled in the segmentation masks.

Training setup. MUNIT achieves unsupervised multi-modal image-to-image translation by

Decoder	Input Shape→Output Shape	Layer Information
$\boldsymbol{G}_\theta(\mathbf{C})$	(128,64,64)→(128,64,64)	CONV-(O:128,K:7x7,S:1,P:3), IN, Leaky ReLU
	(128,64,64)→(128,32,32)	CONV-(O:128,K:4x4,S:2,P:1), IN, Leaky ReLU
	(128,32,32)→(128,16,16)	CONV-(O:128,K:4x4,S:2,P:1), IN, Leaky ReLU
	(128,16,16)→(64,32,32)	DECONV-(O:64,K:4x4,S:2,P:1), IN, Leaky ReLU
	(64,32,32)→(32,64,64)	DECONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,64,64)→(16,128,128)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,128,128)→(3,128,128)	CONV-(O:3,K:7x7,S:1,P:3), Tanh
$\boldsymbol{G}_\theta(\mathbf{s})$	(8)→(256)	FC-(O:256)
	(256)→(4096)	FC-(O:4096)
	(4096)→(8192)	FC-(O:8192), Flatten
	(128,8,8)→(64,16,16)	DECONV-(O:64,K:4x4,S:2,P:1), IN, Leaky ReLU
	(64,16,16)→(32,32,32)	DECONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,32,32)→(16,64,64)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,64,64)→(8,128,128)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,128,128)→(3,128,128)	CONV-(O:3,K:7x7,S:1,P:3), Tanh

Table 4.4: IOB decoders design for MUNIT. The notations in the tables are: O: the number of output channels; K: the kernel size; S: the stride size; P: the padding size; FC: fully-connected layer; IN: instance normalisation;

Decoder	Input Shape→Output Shape	Layer Information
$\boldsymbol{G}_\theta(\mathbf{C})$	(8,224,224)→(8,224,224)	CONV-(O:8,K:7x7,S:1,P:3), IN, Leaky ReLU
	(8,224,224)→(16,112,112)	CONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,112,112)→(32,56,56)	CONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,56,56)→(64,28,28)	CONV-(O:64,K:4x4,S:2,P:1), IN, Leaky ReLU
	(64,28,28)→(128,14,14)	CONV-(O:128,K:4x4,S:2,P:1), IN, Leaky ReLU
	(128,14,14)→(64,28,28)	DECONV-(O:64,K:4x4,S:2,P:1), IN, Leaky ReLU
	(64,28,28)→(32,56,56)	DECONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,56,56)→(16,112,112)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,112,112)→(8,224,224)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,224,224)→(1,224,224)	CONV-(O:1,K:7x7,S:1,P:3), Tanh
$\boldsymbol{G}_\theta(\mathbf{s})$	(3)→(256)	FC-(O:256)
	(256)→(4096)	FC-(O:4096)
	(4096)→(25088)	FC-(O:25088), Flatten
	(128,14,14)→(64,28,28)	DECONV-(O:64,K:4x4,S:2,P:1), IN, Leaky ReLU
	(64,28,28)→(32,56,56)	DECONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,56,56)→(16,112,112)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,112,112)→(8,224,224)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,224,224)→(1,224,224)	CONV-(O:1,K:7x7,S:1,P:3), Tanh

Table 4.5: IOB decoders design for SDNet. The notations in the tables are: O: the number of output channels; K: the kernel size; S: the stride size; P: the padding size; FC: fully-connected layer; IN: instance normalisation;

Decoder	Input Shape→Output Shape	Layer Information
$\boldsymbol{G}_\theta(\mathbf{C})$	(3,64,64)→(16,64,64)	CONV-(O:16,K:7x7,S:1,P:3), IN, Leaky ReLU
	(16,64,64)→(32,32,32)	CONV-(O:32,K:4x4,S:2,P:1), IN, Leaky ReLU
	(32,32,32)→(16,64,64)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,64,64)→(8,128,128)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,128,128)→(3,128,128)	CONV-(O:3,K:7x7,S:1,P:3), Tanh
$\boldsymbol{G}_\theta(\mathbf{s})$	(1024)→(1,32,32)	Flatten
	(1,32,32)→(16,64,64)	DECONV-(O:16,K:4x4,S:2,P:1), IN, Leaky ReLU
	(16,64,64)→(8,128,128)	DECONV-(O:8,K:4x4,S:2,P:1), IN, Leaky ReLU
	(8,128,128)→(3,128,128)	CONV-(O:3,K:7x7,S:1,P:3), Tanh

Table 4.6: IOB decoders design for PANet. The notations in the tables are: O: the number of output channels; K: the kernel size; S: the stride size; P: the padding size; FC: fully-connected layer; IN: instance normalisation;

		Learning Bias	**Design Bias**	
Metric	Original Model	w/o Latent Regression (LR)	w/o AdaIN	w/o Instance Normalisation (IN)
$DC(\mathbf{C}, \mathbf{s})$ (↓)	0.44 ±0.06	**0.40** ±0.08	0.43 ±0.01	0.66 ±0.03
$DC(\mathbf{I}, \mathbf{C})$ (↑)	0.57 ±0.07	0.57 ±0.08	0.58 ±0.08	**0.73** ±0.03
$DC(\mathbf{I}, \mathbf{s})$ (↑)	0.70 ±0.02	**0.73** ±0.03	0.56 ±0.03	0.63 ±0.05
$IOB(\mathbf{I}, \mathbf{C})$ (↑)	4.36 ±0.38	4.34 ±0.58	4.85 ±0.10	**5.01** ±0.12
$IOB(\mathbf{I}, \mathbf{s})$ (↑)	1.31 ±0.04	**1.46** ±0.05	1.17 ±0.04	1.28 ±0.06
FID (↓)	73.48 ±8.35	104.51 ±4.21	**52.48** ±5.03	71.4 ±4.86
LPIPS (↑)	0.08 ±0.01	0.09 ±0.01	0.06 ±0.01	**0.10** ±0.01

Table 4.7: Comparative evaluation of MUNIT variants using the proposed metrics. I use FID and LPIPS to measure translation quality and diversity between SYNTHIA [19] and Cityscapes [20] samples. Results are in "mean ±std" format.

minimizing the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{c-rec} + \lambda_3 \mathcal{L}_{s-rec}, \tag{4.4}$$

where $\mathcal{L}_{rec}$ is the image reconstruction loss i.e. the ℓ_1 distance between the input image and the reconstructed image, $\mathcal{L}_{c-rec}$ and $\mathcal{L}_{s-rec}$ are the content and style latent regression losses, and $\lambda_1 = 10$, $\lambda_2 = 1$ are the hyperparameters used by the authors in [32].

Results. Table 4.7 reports the results of the ablations on the SYNTHIA [19] and Cityscapes [20] datasets. Replacing AdaIN (**w/o AdaIN**) with simple concatenation does not affect the level of C-S disentanglement, but it leads to a 0.14 absolute decrease in $IOB(\mathbf{I}, \mathbf{s})$ and $DC(\mathbf{I}, \mathbf{s})$, indicating that the style becomes less informative and less correlated with the input. Here, I observe an information shift to the content (lower $IOB(\mathbf{I}, \mathbf{s})$, higher $IOB(\mathbf{I}, \mathbf{C})$) leading to better translation quality but worse diversity (LPIPS$= 0.06$). I infer that this variant is worse than the original model, which had more balanced quality/diversity scores. By removing the LR learning bias (**w/o LR**), the style becomes more correlated to the input image. If the style distribution is no longer Gaussian, the style has more degrees of freedom to encode non-relevant information, which contributes to higher $IOB(\mathbf{I}, \mathbf{s})$ and higher C-S disentanglement. This ablation leads to a significant translation quality decrease, while contrary to the analysis

in [32], the diversity is not negatively affected. Finally, by removing IN (**w/o IN**) I expect a more entangled content that is encoding also some style information. The expectations are confirmed by the decrease in C-S disentanglement ($DC(\mathbf{C}, \mathbf{s}) = 0.66$), and a more informative content (which is also more correlated to the input image). Interestingly, relaxing the content constraints for a task that does not require a strictly semantic content (such as image segmentation), leads to the best quality/diversity balance. Note that I define the best balance as achieving the highest average ranking in FID and LPIPS (e.g. the "w/o IN" model variant is the 1^{st} in LPIPS and 2^{nd} in FID).

Summary. The experiments reveal a trade-off between the translation quality/diversity and disentanglement in a translation task. The proposed metrics indicate that a partially disentangled C-S space –with a near-Gaussian style latent space– leads to the best quality/diversity performance. For MUNIT this is achieved by removing the IN design bias.

4.6.3 Medical segmentation

In SDNet, content binarization and style Gaussianity are the key representation constraints. I evaluate their effect and those of decoder design on segmentation performance measuring the Dice Score [195, 196] after: **i)** removing content thresholding (w/o Binarization), **ii)** removing style Gaussianity (w/o Kullback-Liebler Divergence (KLD) and LR), and **iii)** considering a new decoder, obtained replacing the FiLM style conditioning with *SPADE* [105]. SPADE is less restrictive, allowing the style to encode more image-related information, such as textures, rather than just intensity.

Data. I use data from the Automatic Cardiac Diagnosis Challenge (ACDC) [8], which contains cardiac cine-MR images acquired from different MR scanners and resolution on 100 patients. Images were resampled to 1.37 mm/pixel resolution and cropped to 224×224 pixels. Manual segmentations are provided for the left ventricular cavity, the myocardium and right ventricle in the end-systolic and end-diastolic cardiac phases. In total there are 1920 images with manual segmentations and 23,530 images with no segmentations.

Training setup. SDNet is trained by minimizing the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{KL} + \lambda_2 \mathcal{L}_{seg} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{z_{rec}}, \tag{4.5}$$

where $\mathcal{L}_{KL}$ is the KL Divergence measured between the sampled and the predicted style vec-

tors, $\mathcal{L}_{rec}$ is the image reconstruction loss i.e. the ℓ_1 distance between the input image and the reconstructed image, $\mathcal{L}_{seg}$ is the segmentation Dice loss, and $\mathcal{L}_{z_{rec}}$ is the LR loss of the style vector. $\lambda_1 = 0.01, \lambda_2 = 10, \lambda_3 = 1$, and $\lambda_4 = 1$ are the hyperparameters used by the authors in [13].

		Learning Bias	**Design Bias**	
Metric	Original Model	w/o KLD and Latent Reg. (LR)	w/o Binarization	SPADE
$DC(\mathbf{C}, \mathbf{s})$ (↓)	0.49 ±0.02	0.64 ±0.03	**0.44** ±0.00	0.52 ±0.01
$DC(\mathbf{I}, \mathbf{C})$ (↑)	0.94 ±0.01	0.94 ±0.01	**0.98** ±0.02	0.93 ±0.01
$DC(\mathbf{I}, \mathbf{s})$ (↑)	0.43 ±0.02	**0.66** ±0.00	0.44 ±0.01	0.45 ±0.01
$IOB(\mathbf{I}, \mathbf{C})$ (↑)	4.71 ±0.26	4.84 ±0.23	**5.89** ±0.22	5.09 ±0.00
$IOB(\mathbf{I}, \mathbf{s})$ (↑)	1.00 ±0.01	1.00 ±0.04	0.98 ±0.04	1.00 ±0.04
Dice (↑)	0.62 ±0.02	0.61 ±0.04	0.63 ±0.04	**0.75** ±0.02

Table 4.8: Comparative evaluation of SDNet variants using the proposed metrics. I use the Dice score to measure semantic segmentation performance on the ACDC [8] dataset with 1.5% annotation masks. Results are in "mean ±std" format.

		Learning Bias	**Design Bias**	
Metric	Original Model	w/o KLD and Latent Regression	w/o Binarization	SPADE
$DC(\mathbf{C}, \mathbf{s})$ (↓)	0.48	0.57	**0.43**	0.59
$DC(\mathbf{I}, \mathbf{C})$ (↑)	0.97	0.95	**0.97**	0.94
$DC(\mathbf{I}, \mathbf{s})$ (↑)	0.44	0.53	0.44	**0.57**
$IOB(\mathbf{I}, \mathbf{C})$ (↑)	5.66	3.86	**6.21**	5.63
$IOB(\mathbf{I}, \mathbf{s})$ (↑)	0.99	0.96	1.00	**1.02**
Dice (↑)	0.82	0.81	0.82	**0.83**

Table 4.9: Comparative evaluation of SDNet [13] variants on the ACDC [8] dataset with 100% annotation masks, using the proposed metrics. The Dice metric is used to measure the performance in terms of semantic segmentation.

Results. Table 4.8 reports the findings on the ACDC [8] dataset. From the results reported in Table 4.9, it can be seen that when using all the available annotations (fully supervised learning), all SDNet variants achieve a similar accuracy, suggesting that strong learning biases, such as supervised segmentation costs, make disentanglement less important. Thus, I consider the

semi-supervised training case with minimal supervision, using only the 1.5% of available labelled data. Overall, the style encodes little information in all SDNet variants, probably because all medical images in ACDC have similar styles (data bias), and reconstructing using an average style is enough to have low $IOB(\mathbf{I}, \mathbf{s})$. However, C-S disentanglement is still important to obtain a good content representation. For example, intermediate levels of disentanglement (**SPADE**) lead to the best segmentation performance. In this variant, disentanglement decreases compared to the original model, as some style information is probably leaked to the content (higher $DC(\mathbf{C}, \mathbf{s})$ and $IOB(\mathbf{I}, \mathbf{C})$). On the other hand, also removing C binarization (**w/o Binarization**) makes content more informative; since the correlation between C and S decreases, I assume that the extra information encoded in C is not part of the style. Lastly, removing the Gaussian prior constraints from the style (**w/o KLD and LR**) leads to the lowest degree of disentanglement as there is no information bottleneck on S, and a slight decrease of the Dice score.

Summary. I find disentanglement to have minimal effect on task performance when training with strong learning signals (e.g. supervised costs). In the semi-supervised setting, a higher (but not full) degree of disentanglement leads to better performance, while the amount of information in C alone is not enough to achieve adequate segmentation performance.

4.6.4 Pose estimation

I consider the original PANet model and four possible variants, relaxing design biases on both C and style, and learning biases. In detail: **i)** I experiment with a different conditioning mechanism to re-entangle S and C, that consists of the use of AdaIN, rather than just multiplying each S vector with a separate C channel (introducing a bias on S, similar to MUNIT). **ii)** I consider the case where, instead of learning a different S for each channel of C, I extract a global S vector, predicted by an MLP (relaxing the tight 1:1 correspondence between C and S channels). **iii)** I also consider the case where each C part is not approximated by a Gaussian prior. Since we cannot use the original decoder to combine C and S, I reintroduce S using AdaIN. **iv)** Finally, I evaluated the effect of the equivariance constraint, by removing it from the cost function.

Data. I use DeepFashion [21], a large-scale dataset with over 800,000 diverse images of people in different poses and clothing, that also has annotations of body joints. I only used full-body images, specifically 32k images for training and 8k images for testing.

		Learning Bias	**Design Bias**		
Metric	Original Model	w/o Equivar.	AdaIN w/o Gaussian	AdaIN	MLP
$DC(\mathbf{C}, \mathbf{s})$ (↓)	0.65 ±0.01	0.76 ±0.08	**0.25** ±0.01	0.36 ±0.02	0.69 ±0.03
$DC(\mathbf{I}, \mathbf{C})$ (↑)	0.59 ±0.01	**0.60** ±0.02	0.53 ±0.01	0.56 ±0.01	0.58 ±0.02
$DC(\mathbf{I}, \mathbf{s})$ (↑)	**0.83** ±0.01	0.82 ±0.01	0.38 ±0.06	0.81 ±0.01	0.82 ±0.03
$IOB(\mathbf{I}, \mathbf{C})$ (↑)	1.50 ±0.08	1.50 ±0.08	**1.53** ±0.06	1.52 ±0.08	1.49 ±0.06
$IOB(\mathbf{I}, \mathbf{s})$ (↑)	1.09 ±0.04	1.13 ±0.06	1.12 ±0.09	1.10 ±0.15	**1.21** ±0.09
SIM (↑)	**0.71** ±0.02	0.47 ±0.04	0.58 ±0.00	0.64 ±0.01	0.68 ±0.01

Table 4.10: Comparative evaluation of PANet variants using the proposed metrics. I use SIM to measure the performance in terms of pose estimation from landmarks on the DeepFashion [21] dataset. Results are in "mean ±std" format.

Training setup. PANet is trained in an unsupervised way with the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{equiv}, \tag{4.6}$$

where $\mathcal{L}_{rec}$ is the reconstruction loss i.e. the ℓ_1 distance between the input image and the reconstructed image. $\mathcal{L}_{equiv}$ is an equivariance cost, that ensures that the content information does not change after applying some style transformations such as changing the colour of the image. Formally, considering $\mathbf{I}$ is the original image and $\mathbf{I}'$ is the transformed image that has a different style compared to $\mathbf{I}$, $\boldsymbol{E}_{\mathbf{C}}(\mathbf{I})$ encodes the content information. Here, $\mathcal{L}_{equiv} = |\boldsymbol{E}_{\mathbf{C}}(\mathbf{I}) - \boldsymbol{E}_{\mathbf{C}}(\mathbf{I}')|_1$. Based on the implementation details presented in [36], I set $\lambda_1 = \lambda_2 = 1$.

Results. Table 4.10 reports results of the ablations on the DeepFashion [21] dataset. I assess model performance using (Similarity or Histogram Intersection) SIM [197] to measure the similarity between the predicted and ground truth landmarks visualized as heatmaps. Whilst the original model is the best to predict landmarks, it only achieves average disentanglement (see $DC(\mathbf{C}, \mathbf{s})$). Using an **AdaIN**-based decoder consistently improves disentanglement as it has a strong inductive bias on the re-entangled representation (see $DC(\mathbf{C}, \mathbf{s})$ for AdaIN, and **AdaIN w/o Gaussian**), but it leads to worse landmark detection – the representation adapts tightly to the strongly-biased decoder, and the content loses transferability to other tasks. Using an **MLP**

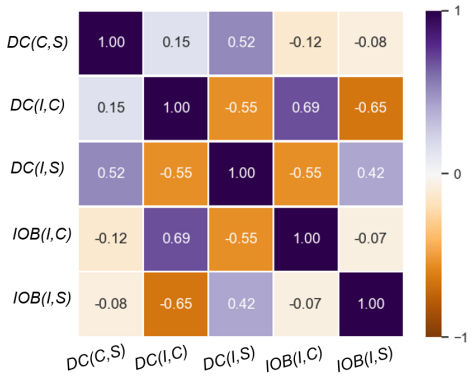


Figure 4.4: Pearson correlation coefficients of the proposed metrics across all models visualized as a heatmap. Values close to 1 and -1 indicate a strong correlation. Figure is taken from [14].

to encode S relaxes the specific conditioning between C and S (a design bias) and reduces disentanglement. In fact, there is an information shift from C to S, as indicated by the higher $IOB(\mathbf{I}, \mathbf{s})$, and I observe a high $DC(\mathbf{C}, \mathbf{s})$. Here, a moderate decrease of disentanglement shows slightly lower task performance. Finally, the equivariance cost is the most important factor for disentanglement; removing it (**w/o Equivariance**) leads to the most entangled representation (high $DC(\mathbf{C}, \mathbf{s})$), and accuracy decrease in landmark detection.

Summary. Overall, learning more entangled representations (higher $DC(\mathbf{C}, \mathbf{s})$ values) leads to better landmark detection. Balance is the key to improve the auxiliary tasks. In PANet, partial disentanglement is achieved by carefully balancing the design biases used to extract the style and to reintroduce it to the content while decoding. Relaxing such biases with AdaIN or MLP makes landmark detection worse.

4.7 Complementary Metrics

As noted in Section 4.3, I report that the proposed metrics are uncorrelated with each other. Here, I present the Pearson correlation computed between disentanglement and performance metrics for each of the investigated models. Intuitively, contrary to the desired low (or no) correlation between disentanglement metrics across all models (see Fig. 4.4), I would expect that the performance metric(s) of each application would be correlated with at least one DC or IOB variant. In fact, this correlation can be exploited to find the "sweet spot" between

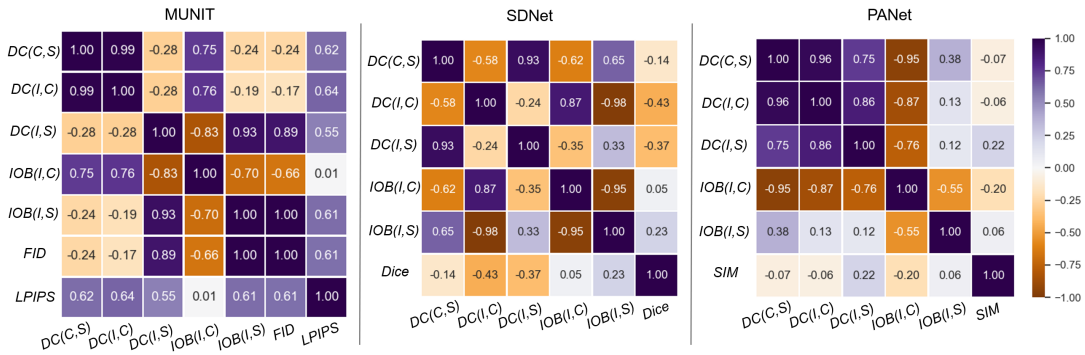


Figure 4.5: Pearson correlation of the proposed metrics across all applications/models visualized as heatmap. Values close to 1 and -1 indicate strong correlation. Figure is taken from [14].

disentanglement and performance. Fig. 4.5 confirms the intuition for all investigated models, highlighting the strong correlation of FID and LPIPS in the MUNIT scenario, which is the only model that utilizes both C and S directly in the main task, e.g. I2I translation.

4.8 Discussion

I now discuss the relationship between C-S disentanglement and inductive biases, task performance, interpretability of the latent representations.

Do biases affect C-S disentanglement? Results in Section 4.6 illustrate that learning and design biases critically affect disentanglement. However, no evaluation can specifically characterize the relative importance of each one, since this depends on the task at hand, as well as the utilized data. In MUNIT, disentanglement is mainly encouraged by the content-related design and learning biases. In fact, IN is key to removing style information from the content, and the model cannot be successfully trained without LR of the content. Disentanglement in SDNet is susceptible to the biases that affect both latent variables. Using a SPADE decoder or removing content thresholding leads to more entanglement, while making the style Gaussian through learning constraints restricts its informativeness and encourages disentanglement. Similarly, PANet disentanglement is affected both by designing the content as Gaussian, and by the equivariance of C and S w.r.t. spatial or intensity transformations, respectively.

What is the relationship between C-S disentanglement and task performance? The results

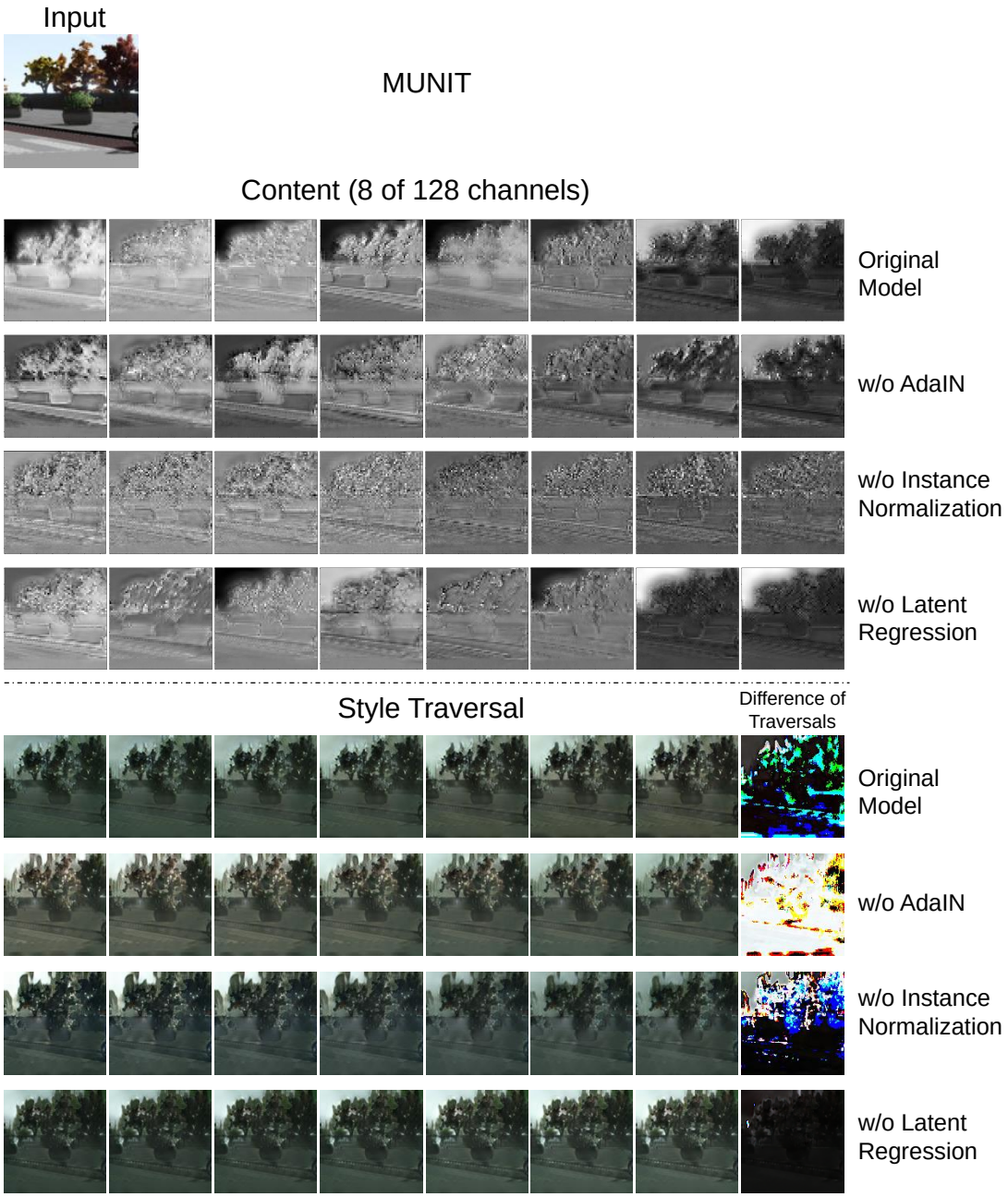


Figure 4.6: MUNIT: Qualitative examples to assess the interpretability of the content and style representations of the investigated model variants for different biases. For each variant, I show 8 channels of the content and 7 indicative style traversals and the difference between the first and last traversal images. The input image is depicted at the top left of the figure. Figure is taken from [14].

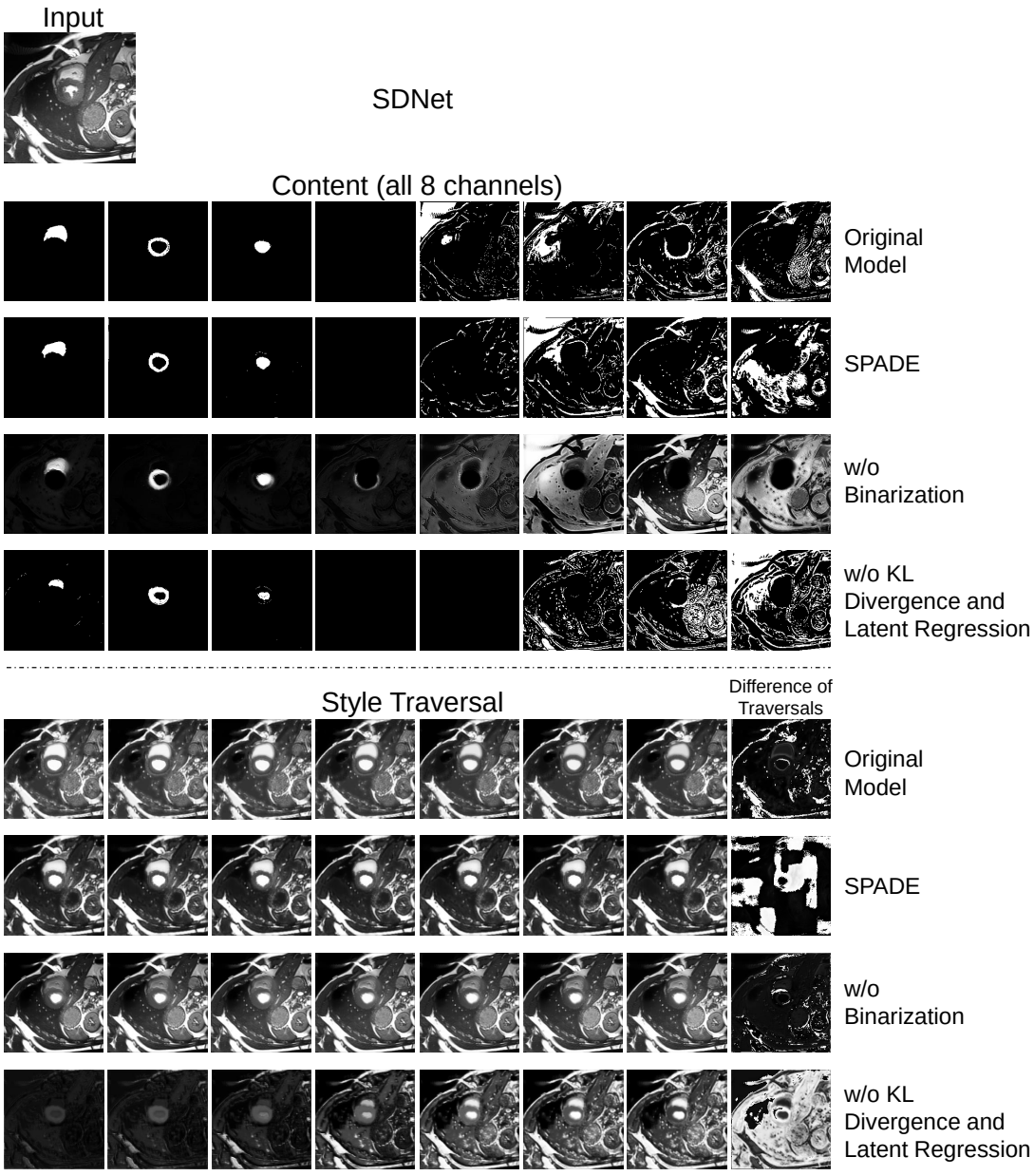


Figure 4.7: SDNet: Qualitative examples to assess the interpretability of the content and style representations of the investigated model variants for different biases. For each variant, I show 8 channels of the content and 7 indicative style traversals and the difference between the first and last traversal images. The input image is depicted at the top left of the figure. Figure is taken from [14].

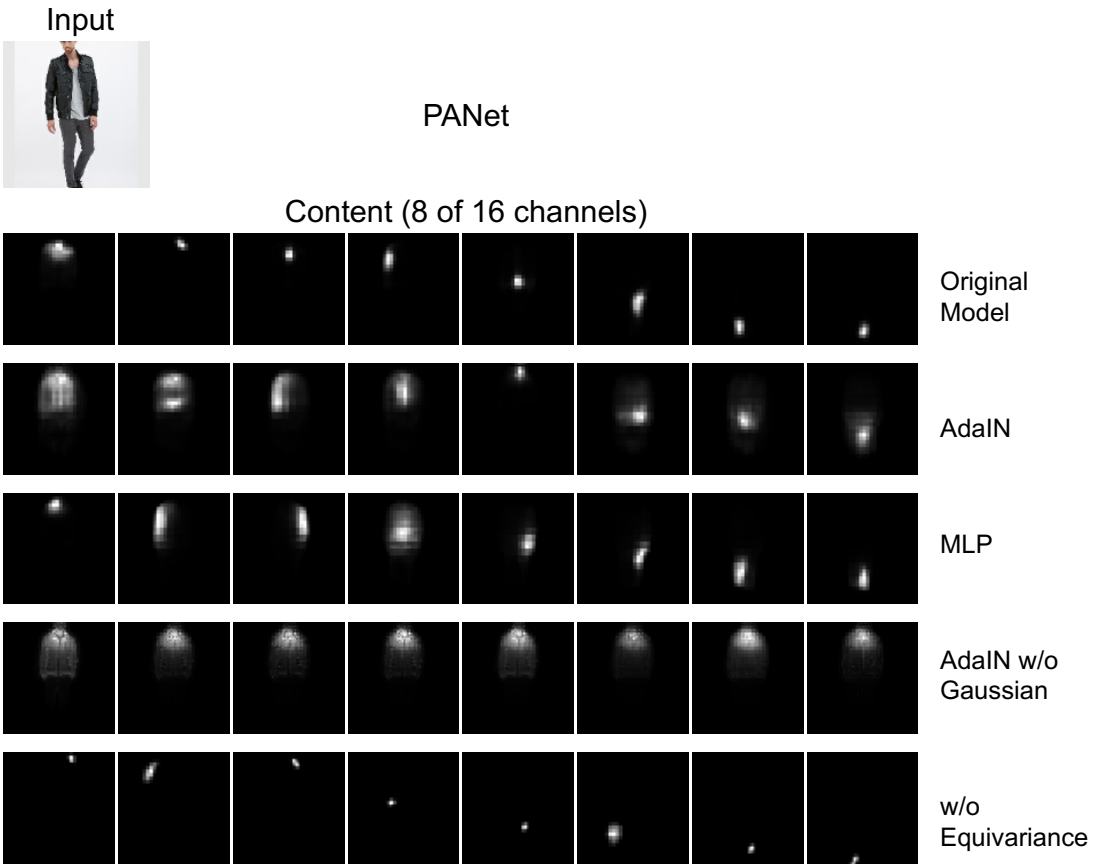


Figure 4.8: PANet: Qualitative examples to assess the interpretability of the content and style representations of the investigated model variants for different biases. For each variant, I show 8 channels of the content. Note that since PANet does not assume a prior distribution on the style, no style are shown. The input image is depicted at the top left of the figure. Figure is taken from [14].

showcase a clear sweet spot between C-S disentanglement and downstream task performance. In particular, I observe that lowering disentanglement by relaxing constraints on the content (e.g. removing IN), but preserving the biases that enforce style priors, such as C-S equivariance, leads to better performance.

Does disentanglement affect content interpretability? Interpretability is hard to quantify without metrics. I visualize the content and style representations in order to reason about their interpretability. I consider the content semantic if distinct objects appear in different channels, whereas the style is semantic when images reconstructed while traversing the style manifold between two points have smooth appearance changes, and are realistic.

I provide visualizations for all model variants. In particular, Figs. 4.6 and 4.7 depict several channels of content, as well as style traversals for different MUNIT and SDNet model variants, respectively. However, Fig. 4.8 presents solely content representations, as PANet does not assume a prior distribution on the style latent vector, thus style traversals are not possible. When interpolating between two style vectors, the originally proposed MUNIT produces realistic images, and smooth appearance changes. Instead, removing the LR constraint affects the image quality. Similarly, the original SDNet presents high image quality and smooth transitions, while removing the content Binarization leads to low intensity (style) diversity.

4.9 Summary

In this chapter, I evaluated the disentanglement between image C and S through experimenting on 3 popular models, and showcased how design and learning biases affect disentanglement and by extension task performance. The findings suggest that whilst content-style disentanglement enables the implementation of certain equivariant tasks, partially (dis)entangled can lead to better performance than fully disentangled ones. Using the findings and the presented metrics will enable the design of better models that achieve the degree of disentanglement that maximizes performance, rather than blindly pursuing very high (or low) disentanglement, which motivates the design of low-rank regularisation in Chapter 5. In addition, the qualitative evaluation of content interpretability inspires how I evaluate compositional equivariance in Chapter 6.

Chapter 5
Disentanglement for Domain-Generalised Medical Image Segmentation

In Chapter 4, I studied how to measure the degree of content-style disentanglement and explored how disentanglement affects task performance. With this knowledge, in this chapter, I focus on deploying disentanglement to improve task performance and to learn models that have better generalisation ability.

5.1 Introduction

I first explore how to augment the training data from multiple source domains such that the model can be trained with more diverse data hence possibly dealing with the domain shifts. Based on our group's previous work SDNet [13], I propose a post-hoc method to randomly mix the latent factors to generate new images to augment the training dataset. I find that this method does not provide guaranteed generalisation ability if the target domain has novel anatomy and modality, which do not exist in the augmented data. Also, the augmentations take advantage of the learnt representation but do not help explicitly in learning better representations. I then explore more advanced approaches to regularise the representations. I propose to combine meta-learning and disentanglement. The proposed approach can capture the latent factors corresponding to certain domain shifts with disentanglement and meta-learning training strategy

This chapter is based on:

- **Liu, X.**, Thermos, S., Chartsias, A., O'Neil, A. and Tsaftaris, S.A., 2020. Disentangled Representations for Domain-generalised Cardiac Segmentation. In International Workshop on Statistical Atlases and Computational Models of the Heart (pp. 187-195). Springer, Cham.
- **Liu, X.**, Thermos, S., O'Neil, A. and Tsaftaris, S.A., 2021. Semi-supervised Meta-learning with Disentanglement for Domain-generalised Medical Image Segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2021.

learns the changes of latent factors across domains. Taking the advantage of unsupervised disentanglement, the method can also be trained with unlabeled data, which addresses the underestimated domain shifts problem of previous meta-learning approaches.

5.1.1 Motivation of the approaches

Despite recent progress in medical image segmentation [8, 198, 199], inference performance on unseen datasets, acquired from distinct scanners or clinical centres, is known to decrease [7, 5]. Such reduction is mainly caused by shifts in data statistics between different clinical centres i.e. *domain shifts* [23], due to variation in patient populations, scanners, and scan acquisition settings [200]. The variation in population impacts the underlying anatomy and pathology due to factors such as gender, age, ethnicity, which may differ for patients in different locations [24, 25, 201]. The variation in scanners and scan acquisition settings impacts the characteristics of the acquired image, such as brightness and contrast [23].

The naive approach to handling domain shift is to acquire and label as many and diverse data as possible, the cost implications and difficulties of which are known to this community. Alternatively one can train a model on source domains to generalise for a target domain with some information on the target domain available i.e. domain adaptation [202] such as cross-site MRI harmonisation [203] to enforce the source and target domains to share similar image-specific characteristics [204]. A more strict alternative is to *not use any* information for the target domain, known as *domain generalisation* [205]. Herein, I focus on this more challenging and more widely applicable approach.

In domain generalisation, the overarching goal is to identify suitable representations that encode information about the task at hand whilst being insensitive to domain-specific information. There are several active research directions aiming to address this goal, including: direct augmentation of the source domain data [23], feature space regularisation [206, 207, 208, 110, 209], alignment of the source domain features or output distributions [210], and learning domain-invariant features with gradient-based meta-learning [211, 212, 213].

As a more advanced approach, gradient-based meta-learning methods have the advantage of not overfitting to dominant source domains which account for the more populous data in the training dataset [212]. Gradient-based meta-learning [205, 214] exploits an episodic training paradigm [215] by splitting the source domains into meta-train and meta-test sets at each it-

eration. The model is trained to handle domain shift by simulating it during training. By using constraints to implicitly eliminate the information related to the simulated domain shifts, the model can learn to extract domain-invariant features. Previous work introduced different constraints in a fully supervised setting e.g. global class alignment and local sample clustering [212], shape-aware constraints [213] or simply the task objective [205, 216], where [216] extends [205] to medical image segmentation.[1] These approaches do not scale in medical image segmentation as pixel-wise annotation is time-consuming, laborious, and requires expert knowledge. Meanwhile, in a low data regime where centres only provide a few labeled data samples, these methods may only learn to extract domain-invariant features from an under-represented data distribution [23, 216]. In other words, the simulated domain shifts may not well approximate the true domain shifts between source and unseen domains.

5.1.2 Approach overview

I first propose two data augmentation methods, termed Resolution Augmentation (RA) and Factor-based Augmentation (FA) (as in Fig. 5.1), which are combined to improve domain adaption and generalisation, thus improving the performance of state-of-the-art models in Cardiac Magnetic Resonance (CMR) image segmentation. In particular, I use RA to remove the resolution bias by randomly rescaling the training images within a predetermined resolution range, while FA is used to increase diversity in the labeled data through mixing spatial and imaging factors, which I denote as the *anatomy* and *modality* factors, respectively. To extract these factors for FA, I pre-train the SDNet model introduced in [13] using the original (prior to augmentation) data. Experiments on the diverse dataset from the STACOM 2020 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms challenge) show the superiority of the proposed methods when combined with the U-Net [218] and SDNet models.

The augmentations only take advantage of the learnt representations e.g. mixing the modality and anatomy factors. To regularise the representations to be generalisable during training, I then propose to explicitly disentangle the representations related to domain shifts for meta-learning as I illustrate in Fig. 5.2. Learning these complete and sufficient representations [56] via reconstruction brings the benefit of unsupervised learning, thus we can better simulate the domain shifts by also using unlabeled data from any of the source domains. I consider two

[1] With the exception of [217] which clusters unlabeled data to generate pseudo labels, but unfortunately is not applicable to segmentation.

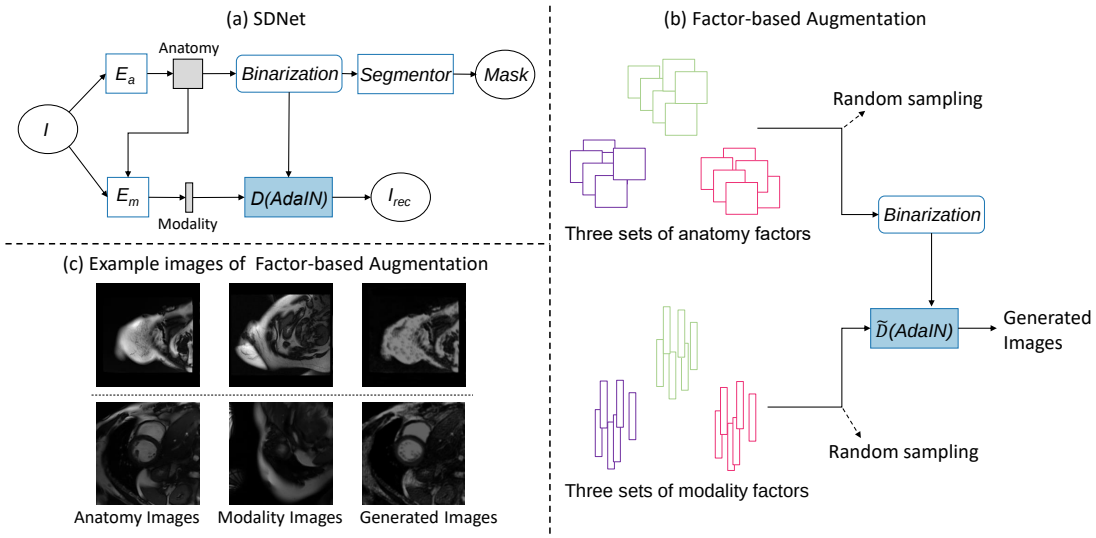


Figure 5.1: (a) SDNet: $\boldsymbol{E}_a$: anatomy encoder, $\boldsymbol{E}_m$: modality encoder, $\boldsymbol{D}(AdaIN)$: AdaIN decoder. I is the input image to the model and I_{rec} is the output image of the AdaIN decoder *i.e.* the reconstructed image. $Mask$ is the predicted segmentation mask for the input image. (b) Illustration of Factor-based Augmentation: $\tilde{\boldsymbol{D}}(AdaIN)$ is a pre-trained AdaIN decoder. (c) Example images produced by Factor-based Augmentation. Anatomy Images provide anatomy factors, Modality Images provide modality factors, and Generated Images are the combination of the anatomy and modality factors. Figure is taken from [15]

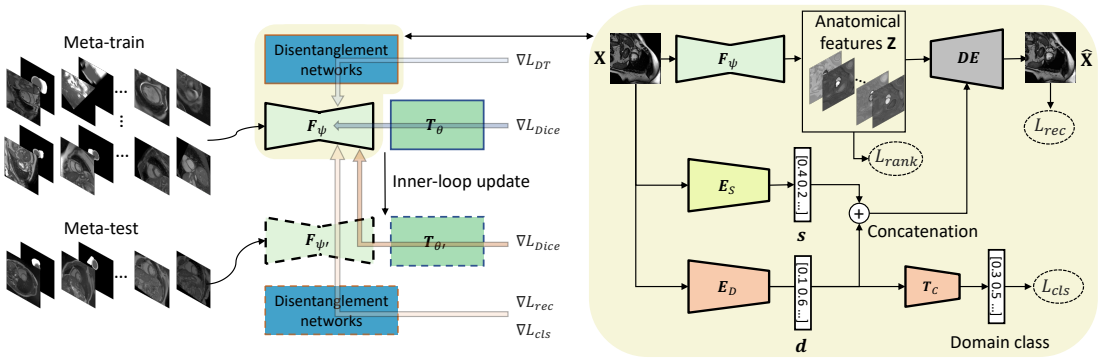


Figure 5.2: At each iteration, the training dataset is split into meta-train and meta-test sets including labeled and unlabeled data. A feature network $\boldsymbol{F}_\psi$ extracts features $\mathbf{Z}$ for a task network $\boldsymbol{T}_\theta$ to predict segmentation masks. The model is trained in a semi-supervised setting, where $\mathcal{L}_{DT}$, $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$ do not require pixel-wise annotation. In the inner-loop update, ψ' and θ' are computed for the meta-test step (see Eq. 5.3). Finally, all the gradients are computed to update $\boldsymbol{F}_\psi$ and $\boldsymbol{T}_\theta$ as in Eq. 5.4. The disentanglement networks decompose image $\mathbf{X}$ to common $\mathbf{s}$ and specific to the domain $\mathbf{d}$ representations to be disentangled with $\mathbf{Z}$ for meta-train and meta-test sets with the constraints ($\mathcal{L}_{DT}$ and $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$). See Section 5.3.1 for loss definitions. Figure is taken from [16]

sources of shifts: one due to scanner and scan acquisition setting variation, and one due to population variation. Because the task is segmentation, we want to be sensitive to changes in anatomy but insensitive to changes in imaging characteristics be it some common across domains or domain-specific. I use spatial (grid-like) features as a representation of anatomy ($\mathbf{Z}$) and two vectors ($\mathbf{s}$,$\mathbf{d}$) to encode common or domain-specific imaging characteristics. I apply specific design and learning biases to disentangle the above. For example, a spatial $\mathbf{Z}$ is equivariant to segmentation and this has been shown to improve performance [32, 13]. I further encourage $\mathbf{Z}$ to be disentangled from $\mathbf{s}$ and $\mathbf{d}$ by exploiting a low-rank regularisation [210]. Gradient-based meta-learning also encourages $\mathbf{Z}$, $\mathbf{s}$, and $\mathbf{d}$ to generalise well to unseen domains whilst at the same time improves (implicitly) their disentanglement.

5.1.3 Contributions

The main contributions of this chapter are summarised as follows:

- I propose two novel augmentation approaches based on disentanglement models.
- I propose the first, to the best of my knowledge, semi-supervised domain-generalisation framework combining meta-learning and disentanglement.
- Use of low-rank regularisation as a learning bias to encourage better disentanglement and hence improved generalisation performance.
- Extensive experiments on cardiac and gray matter datasets show improved performance over several baselines especially for the limited annotated data case.

This chapter is organised as follows. Section 5.2 presents the two proposed augmentation approaches and the corresponding results. Section 5.3 describes the proposed approach that combines meta-learning and disentanglement. Finally, this chapter is concluded in Section 5.4.

5.2 Augmenting the Latent Space for Generalisation

5.2.1 Method

I train the SDNet model and employ RA and FA to generate a more diverse dataset. The model and FA setup is illustrated in Fig. 5.1.

5.2.1.1 Resolution Augmentation (RA):

It is common in MRI data that the imaging resolution (the variation in physical pixel size over image samples) is different for each study due to variation of the scanner parameters. Variation in the imaging resolution can cause the cardiac anatomy to vary significantly in size (i.e. area in pixels), beyond normal anatomical variability.

The training dataset contains subjects scanned by scanners from three vendors i.e. Vendors A, B and C. In Fig. 5.3, I show histograms of the training dataset image resolutions (Section 5.2.2 has more details on the dataset).

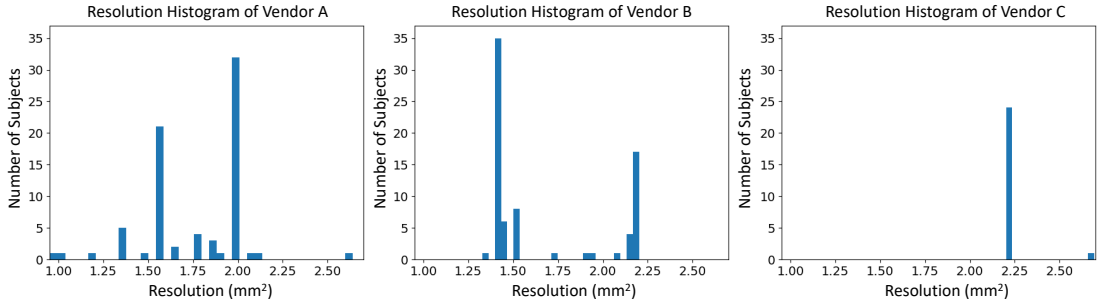


Figure 5.3: Resolution histograms of the M&Ms challenge training data, broken down by vendor (from left to right: Vendors A, B and C).

I observed that the histograms of subjects imaged by scanners from different vendors are distinct from one other i.e. this is a bias with respect to the dataset. To reduce this bias, I propose to augment the training dataset such that the resolutions of subjects are equally distributed from 0.954 mm to 2.692 mm per pixel (the minimum and maximum values observed in the data from the 3 vendors), by rescaling the original image to a random resolution in this range. Finally I center-crop the rescaled image to uniform dimensions of 224×224 pixels.

5.2.1.2 Factor-based Augmentation (FA):

Fig 5.4 illustrates the Factor-based Augmentation method, where a pre-trained SDNet is first used to extract the factors and I mix the factors to generate new images. As shown in Fig. 5.1(a), SDNet decomposes the input image into two latent representations, i.e. anatomy and modality factors, respectively. In particular, the anatomy factor contains spatial information about the input, and the modality factor contains non-spatial information only, namely imaging specific characteristics. Ideally, the anatomy factors would not encode the variation caused by different

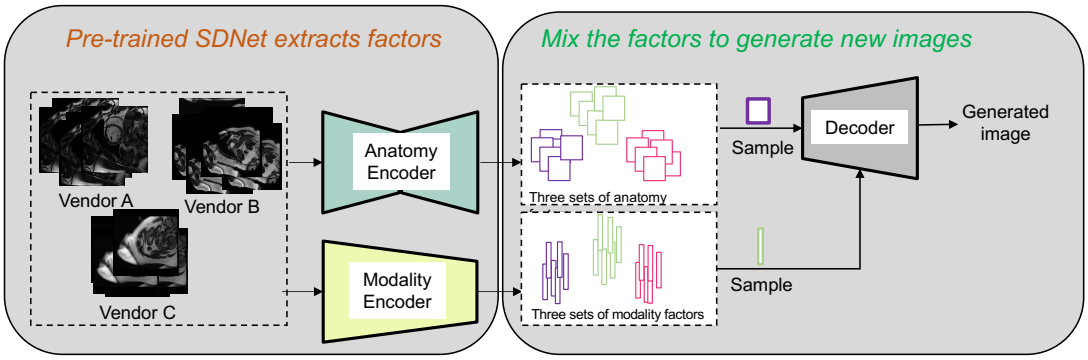


Figure 5.4: For Factor-based Augmentation, I pre-train a SDNet to extract the factors and then I mix the factors to generate new images.

scanners, rather this information would be encoded in the modality factors. Motivated by this, I propose to augment the training dataset by combining different anatomy and modality factors to generate new data.

Considering the three sets of data from Vendors A, B and C, I first pre-train a SDNet model in a semi-supervised manner using the original data. Using this model at inference, I decompose the three sets of data into three sets of anatomy and modality factors. In total, there are 9 possible combinations of the factors resulting in 9 sets of augmented data, where the original training set covers 3 sets and the other 6 sets are novel. As shown in Fig. 5.1(b), I randomly sample an anatomy factor from the three anatomy factor sets and a modality factor from the three modality factor sets. A new image can be generated by processing the two factors with the decoder of the pre-trained SDNet model. By repeating this augmentation process, I generate a larger and more diverse labelled dataset. The segmentation mask of the generated data is the mask of the image providing the anatomy factor, *if* the image is labeled, otherwise the generated data is unlabeled. Some indicative examples of FA are visualized in Fig. 5.1(c).

5.2.1.3 Model Architecture

As depicted in Fig. 5.1(a), the SDNet model consists of 4 modules, namely the anatomy encoder $\boldsymbol{E}_a$, the modality encoder $\boldsymbol{E}_m$, the segmentor, and the AdaIN-based decoder $\boldsymbol{D}(AdaIN)$.

The **anatomy encoder** is realized as a U-Net network that consists of 4 downsampling and upsampling convolutional layers coupled with batch normalisation [103] layers and ReLU [219] non-linearities. The output feature of the anatomy encoder has 8 channels, while the feature

values are thresholded to 0 and 1 by a differentiable rounding operator. By adopting the thresholding constraint and supervision provided by the segmentation masks, the encoded anatomy factor is forced to contain more spatial information about the image.

The **modality encoder** consists of 2 downsampling convolutional layers (4×4 kernel size) that are followed by a global averaging pooling layer, which is used to eliminate the spatial (anatomical) information. The output of the pooling layer is then projected to an 8-dimensional vector (modality factor) using a Multi-Layer Perceptron (MLP) network.

The **segmentor** has 2 convolutional layers using 3×3 kernels, coupled with batch normalisation and ReLU activation layers, as well as a 1×1 convolution followed by a channel-wise softmax. The input to the segmentor is the thresholded anatomy factor. The target is the ground truth segmentation masks when the masks are available. In the learning process, this segmentor encourages the anatomy encoder to encode more spatial information about the image such that the segmentor can learn to predict the corresponding masks more efficiently.

Finally, for the **AdaIN decoder**, I use the AdaIN module as in [102], in order to combine the anatomy and modality representations to reconstruct the image. In particular, the decoder consists of 3 convolutional layers (3×3 kernel size) coupled with adaptive instance normalisation and ReLU activation layers. A final convolutional layer with 7×7 kernels is used for the reconstruction, followed by a hyperbolic tangent activation that normalises the values of the generated image into the [0,1] range. As discussed in [102], the AdaIN decoder normalises the anatomy factor by firstly applying instance normalisation to significantly remove the non-spatial information, then allowing the modality factor to define the new mean and standard derivation of the normalised anatomy factor. In this way, the decoder encourages the anatomy factor to contain spatial information and also force the modality factor to contain non-spatial information. On the other hand, by using AdaIN, the decoder does not simply ignore the modality factor that has a much smaller dimensionality than that of the anatomy factor.

5.2.1.4 Objective function and model training

Apart from the original SDNet objective, I additionally use the focal loss as presented in [220]. Focal loss is widely used in segmentation tasks and helps the model to achieve better accuracy by addressing the class imbalance problem. The augmented overall objective is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{z_{rec}} + \lambda_3 \mathcal{L}_{Dice} + \lambda_4 \mathcal{L}_{focal}, \tag{5.1}$$

where $\mathcal{L}_{rec}$ is the ℓ_1 distance between the input and the reconstructed image. $\mathcal{L}_{z_{rec}}$ denotes the ℓ_1 distance between the encoded modality vector z_{rec} of the original image and the encoded modality vector z'_{rec} of the reconstructed image i.e. latent regression described in Chapter 3.4. $\mathcal{L}_{Dice}$ is the segmentation Dice loss [221]. $\mathcal{L}_{focal}$ is the segmentation focal loss [220] that is defined as:

$$\mathcal{L}_{focal} = -\frac{1}{N} \sum_i \sum_j (1 - \hat{\mathbf{Y}}_i(j))^\gamma \mathbf{Y}_i(j) \log \hat{\mathbf{Y}}_i(j), \tag{5.2}$$

where N denotes the number of pixels of the image or the segmentation mask. $\mathbf{Y}_i(j)$ is the ground truth of segmentation mask for class i at pixel j and $\hat{\mathbf{Y}}_i(j)$ is the prediction. γ is a hyperparameter that is set as 2.

Since I train SDNet in both fully supervised and semi-supervised setups, I set the hyperparameters $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ when training using labeled data, while when training using unlabeled data I set $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = \lambda_4 = 0$.

5.2.2 Experiments

5.2.2.1 Dataset description and preprocessing

I train and validate the proposed method on the M&Ms challenge dataset of 350 subjects. Some subjects have hypertrophic and dilated cardiomyopathies (and some are healthy) but disease labels are not provided. Subjects were scanned in clinical centres in 3 different countries using 4 different magnetic resonance scanner vendors i.e. Vendor A, B, C and D in this section. The M&Ms Challenge training dataset contains 75 labeled subjects scanned using technology from Vendor A, 75 labeled subjects scanned by Vendor B, and 25 unlabeled subjects scanned by Vendor C.I denote subjects scanned by scanners of Vendor A and B as labeled. However, it is notable that only the end diastole and end systole phases are labeled. The M&Ms challenge test dataset contains 200 subjects (50 from each vendor, including the seen 25 unlabeled subjects from Vendor C). From these, 80 subjects (20 from each vendor) are used to validate the model and the rest will be used for final challenge rankings. Subjects scanned by Vendor D were unseen during model training. For each subject, we have 2D cardiac image slice acquisitions captured at multiple phases in the cardiac cycle (including end systole and end diastole). I train

Model	Vendor A			Vendor B			Vendor C			Vendor D		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
U-Net+RA	0.900	0.829	0.811	0.937	0.877	0.907	0.856	0.837	0.852	0.762	0.651	0.503
FS SDNet	0.901	0.837	0.822	0.942	0.877	0.920	0.851	0.826	**0.853**	0.734	0.618	0.474
FS SDNet+RA	0.905	0.846	0.828	**0.945**	0.886	**0.921**	0.855	0.843	0.843	**0.819**	0.749	**0.624**
SS SDNet+RA	**0.909**	**0.854**	**0.841**	**0.945**	**0.887**	0.916	0.843	0.837	0.813	0.811	**0.752**	0.553
SS SDNet+RA+FA	**0.909**	0.846	0.778	0.939	0.882	0.909	**0.863**	**0.847**	0.843	0.812	0.712	0.498

Table 5.1: Evaluation of the 5 models. Average Dice similarity coefficients are reported. Bold numbers denote the best performances across the 5 models. LV: left ventricle, MYO: left ventricular myocardium and RV: right ventricle.

on the 2D images independently because I adopt 2D convolution neural networks in the model. Following [222], I normalise the training data intensity distribution to a Gaussian distribution with a mean of 0 and a standard deviation of 1. Overall, there are 1,738 pairs of images and masks from A and 1,546 pairs of images and masks from B. Apart from these labeled images, there are 47,346 unlabeled images from A, B and C.

5.2.2.2 Model training

Models are trained using the Adam [180] optimizer with an initial learning rate of 0.001. To stabilize the training, I set the new learning rate to 10% of the previous rate when the validation Dice similarity coefficient between the predicted and the ground truth masks does not improve for 2 consecutive epochs. Following the original training setting of SDNet, I set the batch size to 4 and train the model for 50 epochs. All models are implemented in PyTorch [54] and trained using an NVidia 1080 Ti GPU.

5.2.2.3 Results and discussions

To verify the effectiveness of the proposed augmentation methods, I train 5 models for the purpose of ablation: **a)** the U-Net model using samples augmented with RA (U-Net+RA), **b)** the original SDNet model trained in a fully supervised fashion (FS SDNet), **c)** the fully supervised SDNet model using samples augmented with RA (FS SDNet+RA), **d)** the SDNet model trained in a semi-supervised fashion, using samples augmented with RA (SS SDNet+RA), and **e)** the semi-supervised SDNet using samples augmented with both FA and RA (SS SDNet+RA+FA). I train the fully supervised models with labeled samples from vendors A and B, while in the

semi-supervised scenario I use all available data (labeled and unlabeled) for training. Table 5.1 reports the per vendor average Dice scores. Since we are allowed to validate the model a limited number of times, the results are not comprehensive, therefore I will do pairwise analysis below.

5.2.2.4 Does RA help?

By inspection of the **FS SDNet** and **FS SDNet+RA** results, I observe that for Vendor A and B, RA helps to achieve better overall performance compared to the respective baseline and RA substantially improves performance on the unseen Vendor D. Subsequently, for Vendor C, the models have similar performance in the LV class, while FS SDNet+RA achieves the best performance in the MYO class (0.843 Dice score). However, the models using RA do not perform well in the RV class. In the case of Vendor C (no labelled examples), we know that the sample resolution of 24 out of 25 subjects made available at training time from Vendor C is 2.203 mm, and the resolution of 32 out of 75 Vendor A training subjects is around 2.000 mm, thus I argue that the resolution bias for Vendor A and Vendor C are already similar in the original data, and becoming invariant to this bias by rescaling the Vendor A data does not further help the performance on Vendor C.

5.2.2.5 Does FA help?

Inspection of **SS SDNet+RA** and **SS SDNet+RA+FA** results shows that FA has mixed performance. It performs well on Vendor C for the scenario of domain adaptation (target samples available but unlabeled), where Vendor C is one of the vendors providing modality factors. However, FA performs poorly on Vendor D for domain generalisation (unseen target), where Vendor D is not involved in the augmentation process.

5.2.2.6 Best model:

Out of the two augmentation methods, I observe that RA has the most reliable performance, and I choose to submit the **FS SDNet+RA** model for final evaluation in the M&Ms challenge. We can see by comparing **U-Net+RA** with **FS SDNet+RA** that the chosen architecture of SDNet has good generalisation performance compared to a standard U-Net, and this is particularly evident for the unseen vendor D, supporting the choice of the disentangled SDNet architecture even when FA is not employed.

5.2.2.7 Improvements to FA:

Satisfyingly, the FA method yields a benefit for Vendor C (whilst RA did not give a significant benefit to Vendor C). However, FA does not give considerable improvements on domain generalisation. It may be that improvements to the model training would further enhance the benefit of FA. On the other hand, we should also distinguish the effect of simply training on reconstructed images from the effect of FA. This method can be extended to achieve better domain generalisation once I can manipulate the factors realistically to generate out-of-distribution samples.

5.3 Learning to Learn Generalised Disentangled Representations

With the two proposed augmentation methods, I show how to take advantage of the learnt disentangled representations to generate more diverse data. A natural question to answer is if it is possible to regularise the representations such that they are learnt to be more generalisable to address the domain shifts. In this section, I introduce the combination of meta-learning and disentanglement as a solution to the problem of learning more generalisable disentangled representations.

5.3.1 Method

Consider a multi-domain training dataset $\mathcal{D} = \{\mathbf{X}_i^k, \mathbf{Y}_i^k\}_{i=1}^{N_k}, k \in \{1, 2, \cdots, K\}$ that is defined on a joint space $\mathcal{X} \times \mathcal{Y}$, where $\mathbf{X}_i^k$ is the i^{th} training datum from the k^{th} source domain with corresponding ground truth segmentation mask $\mathbf{Y}_i^k$, and N_k denotes the number of training samples in the k^{th} source domain. I aim to learn a model containing a feature network $\boldsymbol{F}_\psi : \mathcal{X} \rightarrow \mathcal{Z}$ to extract the anatomical features $\mathbf{Z}$ and a task network $\boldsymbol{T}_\theta : \mathcal{Z} \rightarrow \mathcal{Y}$ to predict the segmentation masks, where ψ and θ denote the network parameters.

5.3.1.1 Gradient-based meta-learning for domain generalisation

In gradient-based meta-learning for domain generalisation, the domain shift is simulated by training the model on a sequence of episodes [211, 215]. Specifically, the meta-train set $\mathcal{D}_{tr}$ and the meta-test set $\mathcal{D}_{te}$ are constructed by randomly splitting the source domains $\mathcal{D}$ for each iteration of training. Each iteration comprises a meta-train step followed by a meta-test step.

For the meta-train step, the parameters ψ and θ of $\boldsymbol{F}_\psi$ and $\boldsymbol{T}_\theta$ are calculated by optimising the meta-train loss $\mathcal{L}_{meta-train}$ with data from $\mathcal{D}_{tr}$ (inner-loop update), as defined by:

$$(\psi', \theta') = (\psi, \theta) - \alpha \nabla_{\psi,\theta} \mathcal{L}_{meta-train}(\mathcal{D}_{tr}; \psi, \theta), \tag{5.3}$$

where α is the learning rate for the meta-train update step. Typically, $\mathcal{L}_{meta-train}$ is the task objective, e.g. the Dice loss [195] for a segmentation task. This step rewards accurate predictions on the meta-train source domains. For the meta-test step, the meta-test source domains $\mathcal{D}_{te}$ are processed by the updated parameters (ψ', θ') and the model is expected to contain certain properties quantified by the $\mathcal{L}_{meta-test}$ loss. $\mathcal{L}_{meta-test}$ is computed using the updated parameters (ψ', θ'), whilst the gradients are computed towards the original parameters (ψ, θ). The final objective is defined as:

$$\underset{\psi,\theta}{\operatorname{argmin}} \mathcal{L}_{meta-train}(\mathcal{D}_{tr}; \psi, \theta) + \mathcal{L}_{meta-test}(\mathcal{D}_{te}; \psi', \theta'). \tag{5.4}$$

The intuition behind this scheme is that the model should not only perform well on the source domains, but its future updates should also generalise well to unseen domains. Below, I will describe our meta-train and meta-test objectives but first I present how I disentangle representations related to domain shifts.

5.3.1.2 Learning disentangled representations

To model appearance in a single-domain setting, typically a single vector-based variational representation is used [13]. Here, due to our multi-domain setting, inspired by [223, 224], I separately encode domain-specific imaging characteristics as an additional vector-based variational representation. Hence, I aim to learn two independent vector representations, where one ($\mathbf{s}$) captures common imaging characteristics across domains and the other one ($\mathbf{d}$) captures specific imaging characteristics for each domain. In addition, I encode spatial anatomy information in a separate representation $\mathbf{Z}$, which I encourage to be disentangled from $\mathbf{s}$ and $\mathbf{d}$.

In particular, the input image $\mathbf{X}$ is first encoded in a common (appearance) representation $\mathbf{s} = \boldsymbol{E}_S(\mathbf{X})$, and a domain representation $\mathbf{d} = \boldsymbol{E}_D(\mathbf{X})$ that is followed by a shallow domain classifier $\boldsymbol{T}_C(\mathbf{d})$ which predicts the source domain ($\hat{\mathbf{c}}$) label of $\mathbf{X}$. Then, a decoder $\boldsymbol{DE}$ combines the extracted features $\mathbf{Z} = \boldsymbol{F}_\psi(\mathbf{X})$ and the representations $\mathbf{s}$ and $\mathbf{d}$ to reconstruct

the input image, i.e. $\hat{\mathbf{X}} = \boldsymbol{DE}(\mathbf{Z}, \mathbf{s}, \mathbf{d})$. Note that $\boldsymbol{DE}$ combines $\mathbf{Z}$ and $\mathbf{s}, \mathbf{d}$ using adaptive instance normalisation (AdaIN) layers [102]. As shown in [32], AdaIN improves disentanglement and encourages $\mathbf{Z}$ to encode spatially equivariant information, i.e. anatomical information useful for segmentation, and $\mathbf{s}, \mathbf{d}$ to only encode common or domain-specific appearance.

To achieve such "triple" disentanglement I consider several losses: **1)** KL divergences $\mathcal{L}_{KL}(\mathbf{s}, N(0,1)), \mathcal{L}_{KL}(\mathbf{d}, N(0,1))$ to induce a Gaussian $N(0,1)$ prior in $\mathbf{s}$ and $\mathbf{d}$, encouraging the representations to be robust on unseen domains [80]; **2)** Hilbert-Schmidt Independence Criterion (HSIC) loss $\mathcal{L}_{HSIC}(\mathbf{s}, \mathbf{d})$, to force $\mathbf{s}$ and $\mathbf{d}$ to be independent from each other [225]; **3)** a classification loss (ℓ_1 distance) $\mathcal{L}_{cls}(\mathbf{c}, \hat{\mathbf{c}})$ such that the domain representation $\mathbf{d}$ is highly correlated with the domain-specific information [224]; and **4)** a reconstruction loss $\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}})$, defined as the ℓ_1 distance between $\mathbf{X}$ and $\hat{\mathbf{X}}$, to learn representations without supervision [13, 32].

Specifically, considering the kernel function $k : \mathbb{R}^O \times \mathbb{R}^O \rightarrow \mathbb{R}$ where O denotes the vector dimension, the HSIC loss is defined in [225] as:

$$\mathcal{L}_{HSIC}(\mathbf{s}, \mathbf{d}) = (m-1)^{-2}\text{trace}(K_{\mathbf{s}} H K_{\mathbf{d}} H), \tag{5.5}$$

where $m \neq 1$ is the batch size in our case. $K_{\mathbf{s}_{ij}} = k(\mathbf{s}_i, \mathbf{s}_j)$ and $K_{\mathbf{d}_{ij}} = k(\mathbf{d}_i, \mathbf{d}_j)$ are the entries of $K_{\mathbf{s}} \in \mathbb{R}^{m \times m}$ and $K_{\mathbf{d}} \in \mathbb{R}^{m \times m}$. H is the centering matrix $H = I_m - \frac{1}{m} \mathbb{1}_m \mathbb{1}_m^T$. Following [225], we choose the Gaussian kernel $k(\mathbf{s}_i, \mathbf{s}_j) \sim exp(-\frac{1}{2}||\mathbf{s}_i - \mathbf{s}_j||^2/\sigma^2)$, where σ is a hyperparameter that is set as 5.

I further encourage the extracted features $\mathbf{Z}$ to be invariant across the meta-train source domains i.e. invariant to domain shifts and improve disentanglement between $\mathbf{Z}$ and $\mathbf{s}$, $\mathbf{d}$ by applying rank regularisation [210]. Specifically, consider a batch $\{\mathbf{X}^1_{i_1}, \mathbf{X}^2_{i_2}, \cdots, \mathbf{X}^{K_{tr}}_{i_{K_{tr}}}\}$ from K_{tr} meta-train source domains, and K_{tr} features $\{\mathbf{Z}^1_{i_1}, \mathbf{Z}^2_{i_2}, \cdots, \mathbf{Z}^{K_{tr}}_{i_{K_{tr}}}\}$ extracted using the feature network $\boldsymbol{F}_\psi$. By flattening and concatenating these features, we end up with a matrix $\mathbb{Z}$ with dimensions $[C, K_{tr} \times H \times W]$, where C, H, W denote the number of channels, height, and width of $\mathbb{Z}$. Then, by forcing the rank of $\mathbb{Z}$ to be m (i.e. the number of the segmentation classes), $\mathbf{Z}$ is encouraged to encode only globally-shared information across K_{tr} source domains in order to predict the segmentation mask as discussed in [210]. I achieve that by minimising the $(m+1)^{th}$ singular value σ_{m+1} of $\mathbb{Z}$. The rank regularization loss and its gradient can be formulated as:

$$L_{rank} = \sigma_{m+1}, \quad \frac{\partial \sigma_{m+1}}{\partial \mathbb{Z}} = \mathbf{U}\mathbf{V}^T, \tag{5.6}$$

where we perform singular value decomposition (SVD) i.e. $\mathbf{U}, \Sigma, \mathbf{V} = SVD(\mathbb{Z})$ and Σ is the singular value matrix.

Overall, $\mathcal{L}_{DT}$ is defined as:

$$\begin{aligned} \mathcal{L}_{DT} =& \lambda_{rank}\mathcal{L}_{rank}(\mathbf{Z}) + \lambda_{KL}(\mathcal{L}_{KL}(\mathbf{s}, N(0,1)) + \mathcal{L}_{KL}(\mathbf{d}, N(0,1))) \\ &+ \lambda_{rec}\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}}) + \lambda_{HSIC}\mathcal{L}_{HSIC}(\mathbf{s}, \mathbf{d}) + \lambda_{cls}\mathcal{L}_{cls}(\mathbf{c}, \hat{\mathbf{c}}), \end{aligned} \tag{5.7}$$

where $\mathbf{c}$ is the domain label. I adopt hyperparameter values according to our extensive early experiments and discussion from [13, 210] as $\lambda_{rank} = 0.1$, $\lambda_{KL} = 0.1$, $\lambda_{rec} = 1$ and $\lambda_{cls} = 1$. Note that all the losses do not need ground truth masks. The domain class label is available, as we know the centre where the data belong.

5.3.1.3 Meta-train and meta-test objectives

Our *meta-train* objective contains two components:

$$\mathcal{L}_{meta-train} = \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y}, \hat{\mathbf{Y}}) + \mathcal{L}_{DT}, \tag{5.8}$$

where $\lambda_{Dice} = 5$ when labeled data are available.

For the *meta-test* step, the model is expected to: **1)** accurately predict segmentation masks (by applying the task objective), and **2)** disentangle $\mathbf{Z}$ and $\mathbf{s}$, $\mathbf{d}$ to the same level as meta-train sets. A naive strategy for the latter is to use $\mathcal{L}_{DT}$ for meta-test sets. However, as analysed in [226, 213], the meta-test step is unstable to train: the gradients from the meta-test loss are second-order statistics of ψ and θ. Our experiments revealed that including the unsupervised losses $\mathcal{L}_{KL}$ and $\mathcal{L}_{HSIC}$ make training even more unstable (even leading to model collapse). In addition, I use one domain for meta-test in experiments, while $\mathcal{L}_{rank}$ requires multiple domains. According to [31, 14], considering fixed learning and design biases, the level of disentanglement can be proxied by the reconstruction quality (with ground truth image $\mathbf{X}$) and the domain classification accuracy (with ground truth label $\mathbf{c}$). Hence, I adopt as the meta-test loss:

$$\mathcal{L}_{meta-test} = \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y}, \hat{\mathbf{Y}}) + \lambda_{rec}\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}}) + \lambda_{cls}\mathcal{L}_{cls}(\mathbf{c}, \hat{\mathbf{c}}). \tag{5.9}$$

Note that for unlabeled data, $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$ do not need ground truth masks.

5.3.2 Experiments

5.3.2.1 Tasks and datasets

Multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset [5]: The M&Ms challenge dataset contains 320 subjects. Subjects were scanned at 6 clinical centres in 3 different countries using 4 different magnetic resonance scanner vendors (Siemens, Philips, General Electric, and Canon) i.e. domains A, B, C and D. For each subject, only the end systole and end diastole phases are annotated. Voxel resolutions range from $0.85 \times 0.85 \times 10$ mm to $1.45 \times 1.45 \times 9.9$ mm. Domain A contains 95 subjects. Domain B contains 125 subjects. Both domains C and D contain 50 subjects.

Spinal cord gray matter segmentation (SCGM) dataset [7]: The data from SCGM [7] are collected from 4 different medical centres with different MRI systems (Philips Achieva, Siemens Trio, Siemens Skyra) i.e. domains 1, 2, 3 and 4. The voxel resolutions range from $0.25 \times 0.25 \times 2.5$ mm to $0.5 \times 0.5 \times 5$ mm. Each domain has 10 labeled subjects and 10 unlabelled subjects.

5.3.2.2 Baseline models

nnUNet [227]: is a self-adapting framework based on 2D and 3D U-Nets [218] which does not specifically target domain generalisation. Given a labelled training dataset, nnUNet automatically adapts its model design and hyperparameters to obtain optimal performance. In the M&Ms challenge, methods based on nnUNet achieved the top performance [5].

SDNet+Aug. [15]: disentangles the input image to a spatial anatomy and a non-spatial modality factors. Here I use intensity- and resolution- augmented data in a semi-supervised setting. Compared to our method, "SDNet+Aug." only poses disentanglement to the latent features without meta-learning.

LDDG [210]: is the latest state-of-the-art model for domain-generalised medical image analysis. It also uses a rank loss and when applied in a fully supervised setting, LDDG achieved the best generalisation performance on SCGM.

SAML [213]: is another gradient-based meta-learning approach. SAML proposed to enforce the compactness and smoothness properties of segmentation masks across meta-train and meta-

test sets in a fully supervised setting.

5.3.2.3 Implementation details

Models are trained using the Adam optimiser [180] with a learning rate of $2e^{-5}$ for 50K iterations using batch size 4. Images are cropped to 224×244 for M&Ms and 144×144 for SCGM. $\boldsymbol{F}_{\psi}$ is a 2D UNet [218] to extract $\mathbf{Z}$ features with 8 channels of same height and width as input image. I follow the designs of SDNet [13] for $\boldsymbol{E}_S$, $\boldsymbol{T}_{\theta}$ and $\boldsymbol{DE}$. $\boldsymbol{E}_D$ has the same architecture as $\boldsymbol{E}_S$. Both $\mathbf{s}$ and $\mathbf{d}$ have 8 dimensions. $\boldsymbol{T}_C$ is a single fully-connected layer. I use AdaIN module (as described in Chapter 3) in the decoder $\boldsymbol{DE}$ to combine $\mathbf{Z}$ and $\mathbf{s}, \mathbf{d}$.

All models are implemented in PyTorch [54] and are trained using an NVidia 2080 Ti GPU. In the semi-supervised setting, I use specific percentages of the subjects as labeled data and the rest as unlabeled data. I use Dice (%) and Hausdorff Distance [228] as the evaluation metrics.

5.3.2.4 Qualitative results

I show the qualitative results, i.e. predicted segmentation masks, in Fig. 5.5. When training the baseline models with less labeled data, the performance drops significantly. In contrast, our model can produce satisfactory masks in every case.

5.3.2.5 Quantitative results and discussion

Tables 5.2, 5.3, 5.4 and 5.5 show that the proposed method consistently achieves the best generalisation performance on cardiac and gray matter segmentation. Particularly in the low data regime I improve Dice by $\approx 5\%$ on M&Ms and $\approx 3\%$ on SCGM compared to the best performing baseline. For 100% annotations in M&Ms, our model still outperforms the baselines.

M&Ms: Compared to "SDNet+Aug." which can also use (due to disentanglement) unlabeled data, our model performs consistently better. The results agree with the conclusion in [39]: without specific designs tuned to the tasks, disentanglement can not provide guaranteed generalisation ability. For LDDG and SAML, the generalisation performance significantly drops with small amounts of labeled data. Note that nnUNet adapts the model design per each run/training set. However, adapting the model design for different training data limits the scalability of nnUNet. I also report the Dice (%) and Hausdorff Distance results on the cases of giving 100%

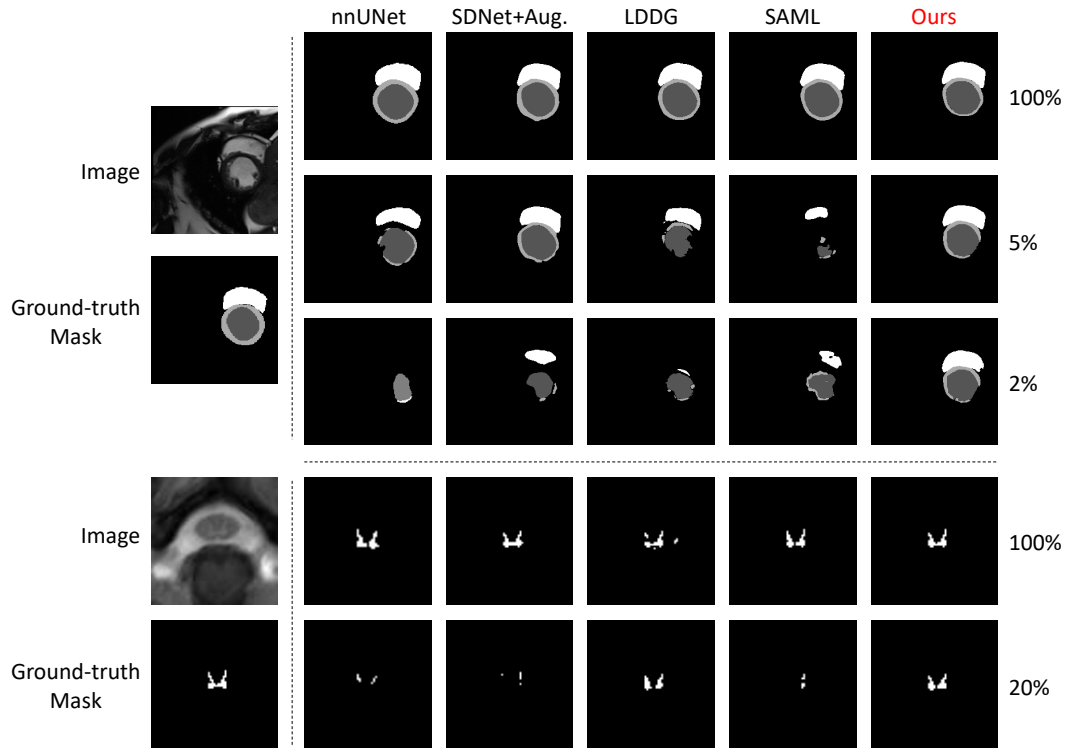


Figure 5.5: I show the example images and predicted segmentation masks of each model for different cases.

Source		**Target**	**nnUNet**	**SDNet+Aug.**	**LDDG**	**SAML**	**Ours**
2%	B,C,D	A	52.87_{19}	54.48_{18}	59.47_{12}	56.31_{13}	$\mathbf{66.01_{12}}$
	A,C,D	B	64.63_{17}	67.81_{14}	56.16_{14}	56.32_{15}	$\mathbf{72.72_{10}}$
	A,B,D	C	72.97_{14}	76.46_{12}	68.21_{11}	$75.70_{8.7}$	$\mathbf{77.54_{10}}$
	A,B,C	D	73.27_{11}	74.35_{11}	68.56_{10}	$69.94_{9.8}$	$\mathbf{75.14_{8.4}}$
	Average		$65.94_{8.3}$	$68.28_{8.6}$	$63.16_{5.4}$	$64.57_{8.5}$	$\mathbf{72.85_{4.3}}$
5%	B,C,D	A	65.30_{17}	71.21_{13}	$66.22_{9.1}$	67.11_{10}	$\mathbf{72.40_{12}}$
	A,C,D	B	79.73_{10}	77.31_{10}	$69.49_{8.3}$	$76.35_{7.9}$	$\mathbf{80.30_{9.1}}$
	A,B,D	C	78.06_{11}	$81.40_{8.0}$	$73.40_{9.8}$	$77.43_{8.3}$	$\mathbf{82.51_{6.6}}$
	A,B,C	D	$81.25_{8.3}$	$79.95_{7.8}$	$75.66_{8.5}$	$78.64_{5.8}$	$\mathbf{83.77_{5.1}}$
	Average		$76.09_{6.3}$	$77.47_{3.9}$	$71.29_{3.6}$	$74.88_{4.6}$	$\mathbf{79.75_{4.4}}$
100%	B,C,D	A	80.84_{11}	$81.50_{7.7}$	$82.62_{6.3}$	$81.33_{7.2}$	$\mathbf{83.21_{7.4}}$
	A,C,D	B	$\mathbf{86.76_{5.8}}$	$85.04_{6.1}$	$85.68_{5.7}$	$84.15_{5.9}$	$86.53_{5.3}$
	A,B,D	C	$84.92_{7.1}$	$85.64_{6.5}$	$86.49_{6.3}$	$84.52_{6.2}$	$\mathbf{87.22_{6.1}}$
	A,B,C	D	$86.94_{5.9}$	$84.96_{5.2}$	$86.73_{6.1}$	$83.96_{5.9}$	$\mathbf{87.16_{4.9}}$
	Average		$84.87_{2.5}$	$84.29_{1.6}$	$85.38_{1.6}$	$83.49_{1.3}$	$\mathbf{86.03_{1.7}}$

Table 5.2: Dice (%) results and the standard deviations on M&Ms dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 2% or 5% of labeled data from source domains. The other models are trained by 2% or 5% labeled data only. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
2%	B,C,D	A	$26.48_{7.5}$	$24.69_{7.0}$	$25.56_{5.9}$	$25.57_{5.7}$	$\mathbf{23.55_{6.5}}$
	A,C,D	B	$23.11_{6.8}$	$21.84_{6.2}$	$25.44_{5.2}$	$24.91_{5.5}$	$\mathbf{19.95_{6.3}}$
	A,B,D	C	$16.75_{4.6}$	$16.57_{4.2}$	$18.98_{3.9}$	$16.46_{3.5}$	$\mathbf{16.29_{4.0}}$
	A,B,C	D	$17.51_{4.9}$	$17.57_{4.1}$	$18.08_{3.8}$	$17.94_{3.8}$	$\mathbf{17.48_{4.7}}$
	Average		$20.96_{4.0}$	$20.17_{3.3}$	$22.02_{3.5}$	$21.22_{4.1}$	$\mathbf{19.32_{2.8}}$
5%	B,C,D	A	$23.04_{6.7}$	$22.84_{6.3}$	$23.35_{5.7}$	$23.10_{5.9}$	$\mathbf{22.55_{6.6}}$
	A,C,D	B	$\mathbf{18.18_{4.7}}$	$20.26_{5.5}$	$20.56_{4.7}$	$18.97_{4.9}$	$19.37_{6.4}$
	A,B,D	C	$16.44_{4.2}$	$16.22_{3.9}$	$17.14_{3.3}$	$16.29_{3.2}$	$\mathbf{15.77_{3.8}}$
	A,B,C	D	$15.24_{4.2}$	$15.15_{3.3}$	$15.80_{3.2}$	$15.58_{3.2}$	$\mathbf{14.24_{2.8}}$
	Average		$18.22_{3.0}$	$18.62_{3.1}$	$19.21_{3.0}$	$18.49_{2.9}$	$\mathbf{17.98_{3.2}}$
100%	B,C,D	A	$17.86_{5.5}$	$17.39_{4.5}$	$17.48_{4.1}$	$17.70_{4.2}$	$\mathbf{17.28_{3.9}}$
	A,C,D	B	$\mathbf{14.82_{3.4}}$	$15.55_{3.7}$	$15.42_{3.4}$	$16.05_{3.7}$	$14.99_{3.6}$
	A,B,D	C	$13.72_{3.3}$	$13.67_{3.0}$	$13.52_{2.8}$	$14.21_{3.3}$	$\mathbf{13.11_{2.8}}$
	A,B,C	D	$12.81_{3.4}$	$13.64_{2.9}$	$13.11_{3.0}$	$14.12_{2.8}$	$\mathbf{12.72_{2.6}}$
	Average		$14.80_{1.9}$	$15.06_{1.6}$	$14.88_{1.7}$	$15.52_{1.5}$	$\mathbf{14.53_{1.8}}$

Table 5.3: Hausdorff distance results and the standard deviations on M&Ms dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 2% or 5% or 100% of labeled data from source domains. The other models are trained by 2% or 5% or 100% labeled data only. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
20%	2,3,4	1	59.07_{21}	83.07_{16}	$77.71_{9.1}$	78.71_{25}	$\mathbf{87.45_{6.3}}$
	1,3,4	2	69.94_{12}	$80.01_{5.2}$	44.08_{12}	75.58_{12}	$\mathbf{81.05_{5.2}}$
	1,2,4	3	$60.25_{7.2}$	58.57_{10}	$48.04_{5.5}$	$54.36_{7.6}$	$\mathbf{61.85_{7.3}}$
	1,2,3	4	$70.13_{4.3}$	$85.27_{2.2}$	$83.42_{2.7}$	$85.36_{2.8}$	$\mathbf{87.96_{2.1}}$
	Average		$64.85_{5.2}$	76.73_{11}	63.31_{17}	73.50_{12}	$\mathbf{79.58_{11}}$
100%	2,3,4	1	$75.27_{8.3}$	$\mathbf{90.25_{4.5}}$	$88.21_{4.9}$	$90.22_{5.6}$	$90.01_{4.9}$
	1,3,4	2	$76.32_{2.9}$	$84.13_{4.2}$	$83.76_{3.1}$	$\mathbf{86.65_{3.5}}$	$85.48_{2.3}$
	1,2,4	3	$62.59_{6.9}$	62.18_{10}	$56.11_{9.3}$	$58.27_{9.4}$	$\mathbf{64.23_{9.7}}$
	1,2,3	4	$71.87_{2.5}$	$88.93_{1.9}$	$89.08_{2.7}$	$88.66_{2.6}$	$\mathbf{89.26_{2.5}}$
	Average		$71.51_{5.4}$	81.37_{11}	79.29_{13}	80.95_{13}	$\mathbf{82.25_{11}}$

Table 5.4: Dice (%) results and the standard deviations on SCGM dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 20% or 100% of labeled data from source domains. The other models are trained by 20% or 100% of labeled data only. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	Ours
20%	2,3,4	1	$3.09_{0.25}$	$1.52_{0.33}$	$1.75_{0.26}$	$1.53_{0.38}$	$\mathbf{1.50_{0.30}}$
	1,3,4	2	$3.16_{0.09}$	$1.97_{0.16}$	$2.73_{0.33}$	$2.07_{0.35}$	$\mathbf{1.91_{0.16}}$
	1,2,4	3	$3.38_{0.27}$	$2.45_{0.27}$	$2.67_{0.25}$	$2.52_{0.24}$	$\mathbf{2.23_{0.23}}$
	1,2,3	4	$4.31_{0.14}$	$2.34_{0.21}$	$2.37_{0.14}$	$2.30_{0.18}$	$\mathbf{2.22_{0.13}}$
	Average		$3.49_{0.49}$	$2.07_{0.36}$	$2.38_{0.39}$	$2.11_{0.37}$	$\mathbf{1.97_{0.30}}$
100%	2,3,4	1	$3.26_{0.21}$	$\mathbf{1.37_{0.25}}$	$1.50_{0.23}$	$1.43_{0.36}$	$1.43_{0.29}$
	1,3,4	2	$3.19_{0.09}$	$1.88_{0.16}$	$2.19_{0.19}$	$\mathbf{1.80_{0.19}}$	$1.81_{0.15}$
	1,2,4	3	$3.37_{0.27}$	$2.34_{0.24}$	$2.64_{0.28}$	$2.43_{0.33}$	$\mathbf{2.23_{0.32}}$
	1,2,3	4	$4.30_{0.15}$	$2.13_{0.17}$	$2.12_{0.15}$	$2.15_{0.15}$	$\mathbf{2.11_{0.13}}$
	Average		$3.53_{0.45}$	$1.93_{0.36}$	$2.11_{0.41}$	$1.95_{0.38}$	$\mathbf{1.92_{0.31}}$

Table 5.5: Hausdorff distance results and the standard deviations on SCGM dataset. For "SDNet+Aug." and our method, the training data contain all the unlabeled data and 20% or 100% of labeled data from source domains. The other models are trained by 20% or 100% of labeled data only. Bold numbers denote the best performance.

Source		Target A	Target B	Target C	Target D
100%	Trained on B	78.82_{11}	$94.58_{4.3}$	$83.60_{8.0}$	$85.81_{6.8}$
	Trained on D	80.03_{10}	$84.35_{7.0}$	$84.01_{8.9}$	$94.74_{5.4}$

Table 5.6: Dice (%) results and the standard deviations on M&Ms dataset.

labeled data for each model. For M&Ms, the unlabeled data (of labeled subjects) from phases between end-systole and end diastole phases still gives the proposed model slightly better performance i.e. 0.65% of improvement compared with the best baseline model.

SCGM: I obtain consistent improvements also on SCGM, demonstrating application in other organs. The model benefits from the additional 10 unlabeled subjects of each domain leading to better performance overall.

5.3.2.6 Analysis of M&Ms data

To explore why nnUNet can outperform other models on cases of A,C,D to B and A,B,C to D, I train nnUNet models with 100% labeled data only from domain B or domain D. Then I test the models with data from domain A, B, C and D. As shown in Table 5.6, the model trained on domain B can achieve 85.81% Dice on domain D (highest Dice compared to A and C). Also, the model trained on domain D can achieve 84.35% Dice on domain B (highest Dice compared to A and C). Hence, nnUNet possibly overfits the source domains e.g. B, to achieve the best performance on domains (e.g. D) similar to B. However, other methods have to generalise to distinct domains i.e. A and C, which causes slightly worse performance on the domains similar to source domains.

5.3.2.7 Ablation study

Here I conduct ablations on key losses crucial to disentanglement and the extraction of good anatomical features for good generalisation performance. I omit ablations on the KL losses as [80, 224] showcase that variational encoding helps to learn robust vector representation for better generalisation. To illustrate that $\mathcal{L}_{rank}$ helps to disentangle $\mathbf{Z}$ to $(\mathbf{s}, \mathbf{d})$, and improves performance, I use Distance Correlation (DC) [14] (described in Chapter 4) to measure disentanglement (lower DC means a higher level of disentanglement). For M&Ms 5% cases, without $\mathcal{L}_{rank}$, the average DC on the test dataset between $\mathbf{Z}$ and $(\mathbf{s}, \mathbf{d})$ is 0.22 (an increase

compared to 0.19 with $\mathcal{L}_{rank}$), and the average Dice is 78.54% (a decrease compared to 79.75% with $\mathcal{L}_{rank}$). I also ablate $\mathcal{L}_{cls}$ and $\mathcal{L}_{HSIC}$. The proposed model on M&Ms 5% cases had an average Dice 79.75% but without $\mathcal{L}_{cls}$, average Dice drops to 77.45% and without $\mathcal{L}_{HSIC}$, average Dice drops to 77.86%.

5.4 Summary

In this chapter, two data augmentation methods were first proposed to address the domain adaptation and generalisation problems in the field of CMR image segmentation. In particular, a geometry-related augmentation method was introduced, which aims to remove the scale and resolution bias of the original data. Further, the second proposed augmentation method aims bridge the gap between populations and data captured by different scanners. To achieve this, the original data is projected onto a disentangled latent space and generates new samples by combining disentangled factors from different domains. The presented experimental results showcase the contribution of the geometry-based method to CMR image segmentation through improving the domain generalisation, while also demonstrating the contribution of the disentangled factors mixing method to the domain adaptation. Then, I have presented a novel semi-supervised meta-learning framework for domain generalisation. Using disentanglement, the proposed approach models domain shifts, and thanks to the reconstruction approach to disentanglement, the model can be trained also with unlabeled data. By applying the designed constraints (including the low-rank regularisation) to the gradient-based meta-learning approach, the model extracts robust anatomical features useful for predicting segmentation masks in a semi-supervised manner. Extensive quantitative results, especially when insufficient annotated data are available, indicate remarkable improvements compared to previous state-of-the-art approaches. It is notable that both methods are considering enforcing independence in the latent space to learn disentangled representations. In practice, the generative factors are not necessarily always independent of each other. In this case, enforcing independence does not help in learning the representations that approximate well the true generative factors, leading to a limitation on generalising to new domains. In the next chapter, I will discuss how to consider compositionality as a proxy for learning better representations.

Chapter 6
Compositional Representation Learning

In Chapter 5, I presented the methods for generalisation based on learning disentangled representations. As I discussed in Chapter 3 and Chapter 4, most of the disentanglement methods enforce independence or uncorrelation between the latent representations. The SDNet [13] backbone I used for the data augmentations and the proposed meta-learning disentanglement approach are not excluded from the independent representation learning family in Chapter 5. In this chapter, I study how compositionality can be taken into account as a better prior to learn more generalisable and interpretable representations.

6.1 Introduction

The real world has way more complex generative factors that are not trivially independent of each other. For example, as discussed in [43], there exhibits a strong positive correlation in observed data between foot length and body height. When learning the representations for the foot length and body height, forcing them to be independent does not approximate well the reality. In this chapter, I will study using compositionality as a prior to learning the representations i.e. compositional representation learning. In this case, when generating the images, I compose the representations, where the relationship between representations is modeled with the composing operations that are learnt from the data. Overall, compositionality is a better inductive bias beyond independence for learning generalisable and interpretable representations. More importantly, extensive annotations are not necessary for learning compositional representations.

This chapter is based on:

- **Liu, X.**, Thermos, S., Sanchez, P., O'Neil, A. and Tsaftaris, S.A., 2022. vMFNet: Compositionality Meets Domain-generalised Segmentation. In International Conference on Medical Image Computing and Computer Assisted Intervention 2022.
- **Liu, X.**, Sanchez, P., Thermos, S., O'Neil, A. and Tsaftaris, S.A., 2023. Compositionally Equivariant Representation Learning. IEEE Transactions on Medical Imaging (under review).

6.1.1 Motivation of the approach

When a large amount of labelled training data are available, deep learning techniques have demonstrated remarkable accuracy in medical image analysis tasks like diagnosis and segmentation [8]. However, by contrast, humans are able to learn quickly with only limited supervision, and their recognition is not only fast but also robust and easily generalisable [153, 158]. For instance, clinical experts tend to remember configurations (components) of human anatomical structures from multiple medical images they have seen. When searching for anatomy of interest in new images, they use these configurations to locate and identify the anatomy in the image. This compositionality has been shown to enhance the robustness and interpretability in computer vision tasks [153, 157, 149] but has received limited attention in medical applications.

6.1.2 Approach overview

In this chapter, I investigate the application of compositionality to learn good representations in the medical field. Drawing inspiration from Compositional Networks [149]. I model the compositional representations of human anatomy as learnable von-Mises-Fisher (vMF) kernels. Note that vMF kernels are similar to prototypes in [229]. However, prototypes are often calculated as the mean of feature vectors for each class using the ground truth masks, while vMF kernels are learned as the cluster centres of the feature vectors. Considering that medical images are first processed by deep models into deep features, I transform the features into vMF activations that determine the extent to which each kernel is activated at each position. Without any other constraints, the compositional representations do not carry meaningful information that corresponds to the generative factors. I claim that each generative factor is compositional (e.g. the patterns of human anatomy) and also equivariant to the task, i.e. compositionally equivariant. To approximate well the generative factors, I consider different settings i.e. un-, weakly-, and semi-supervised settings and different learning biases that enforce the representations to be more compositionally equivariant. An approach overview is included in Fig. 6.1.

To evaluate the level of compositional equivariance, I measure the interpretability and generalisation ability of the representations. I first qualitatively evaluate the interpretability of the activations of each representation for different settings. As expected, I observe that stronger learning biases (e.g. weak supervision or some supervision) lead to better interpretability. Then, I consider the task of semi-supervised domain generalisation [16, 12, 205] on medical image segmentation and compare my methods with several strong baselines. Extensive quan-

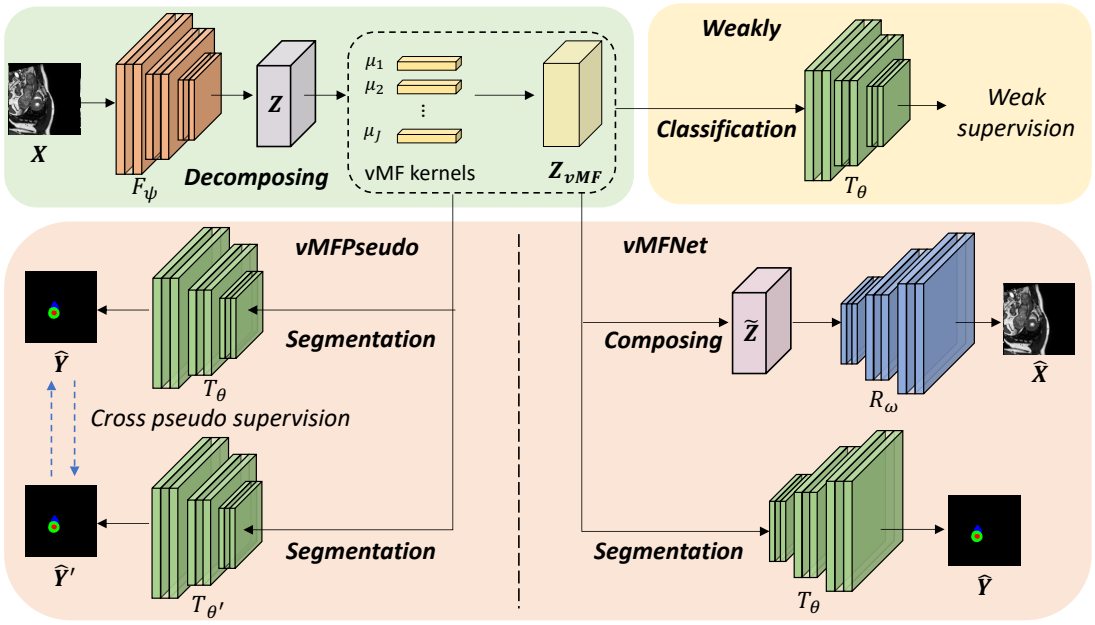


Figure 6.1: The overview of compositionally equivariant representation learning. After decomposing the image features into compositional kernels, different design and learning biases are considered under different settings.

titative results on the multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset [5] and spinal cord gray matter segmentation (SCGM) dataset [7] show that the compositionally equivariant representations have superior generalisation ability, achieving state-of-the-art performance.

This chapter is based on vMFNet that I published at the MICCAI conference [12]. Compared to vMFNet, I propose the compositional equivariance theory. I consider more learning settings as well as more design and learning biases to learn the compositional representations. vMFNet is only one out of the four methods. Moreover, I conduct more experiments, especially on the proposed semi-supervised setting with pseudo supervision on the domain generalisation setting, where better results are observed for some cases compared to vMFNet. I believe that this chapter demonstrates more comprehensively the benefits and potential of the application of compositionality in the medical domain.

6.1.3 Contributions

Overall, the contributions are the following:

- I propose the compositionality theory and propose that the generative factors satisfy the compositional equivariance property.
- By modelling the compositional representations with the vMF kernels, I study different settings and different learning biases that can be used to learn compositional equivariant representations.
- I consider the interpretability and generalisation ability of the learnt representations as the measurement of compositional equivariance.
- I propose a reconstruction module to compose the vMF kernels with the vMF likelihoods to facilitate reconstruction of the input image, which allows the model to be trained also with unlabeled data.
- I apply the proposed method to two settings: semi-supervised domain generalisation and test-time domain generalisation.
- I perform extensive experiments on two medical datasets and compare our methods with several strong baselines.
- I show that different learning biases can help to achieve different levels of compositional equivariance with extensive qualitative and quantitative results.

6.2 Related work

6.2.1 Compositionality

Compositionality has been mostly utilized in robust image classification [230, 153, 149] and recently in compositional image synthesis [158, 159]. Among these works, Compositional Networks [149] — designed originally for robust classification under object occlusion — can be easily adapted to pixel-wise tasks as they learn spatial and interpretable vMF activations. Previous research has combined vMF kernels and activations [149] for object localisation [160] and, recently, for nuclei segmentation (with bounding box supervision) in a weakly supervised manner [161]. In the proposed approach, I model compositional representations using vMF kernels. By incorporating more learning biases that constrain the kernels, we can assign information about each generative factor more specifically to each kernel, resulting in compositional

equivariance. Using unlabelled data, I learn vMF kernels and activations in a semi-supervised manner for domain-generalised medical image segmentation.

6.2.2 Domain generalisation

Many solutions are proposed to address the domain generalisation problem in medical image analysis. Various methods have been used, such as augmentation of the source domain data [23, 231], regularisation of the feature space [208, 110], alignment of the source domain features or output distributions [210], design of robust network modules [232], or the use of meta-learning to adapt to possible domain shifts [212, 213, 109, 16]. Most of these approaches are based on fully supervised learning. More recently, a gradient-based meta-learning model was proposed to handle semi-supervised domain generalisation by integrating disentanglement [16]. Another method used a pre-trained ResNet as a backbone feature extractor, augmenting the source data, and leveraging the unlabelled data through pseudo-labelling [233]. The proposed approach aligns image features to the same von-Mises-Fisher distributions to handle domain shifts. In the semi-supervised setting with reconstruction, the reconstruction further enables the model to handle domain generalisation with unlabelled data. For the semi-supervised setting with pseudo supervision, the pseudo supervision as in [233] enables the model to be trained with unlabelled data and the final prediction is equivalently ensembled from two models.

6.3 Method

I denote x as a scalar, $\mathbf{x}$ as a vector and $\mathbf{X}$ as a tensor. Consider a dataset $\mathcal{D} = \{\mathbf{X}_i, \mathbf{Y}_i\}_{i=1}^{N}$ that is defined on a joint space $\mathcal{X} \times \mathcal{Y}$, where $\mathbf{X}_i$ is the i^{th} training datum with corresponding ground truth label $\mathbf{Y}_i$ (e.g. for a segmentation task, $\mathbf{Y}_i$ is the ground truth segmentation mask), and N denotes the number of training samples. I aim to learn a model containing a representation encoding network $\boldsymbol{F}_\psi : \mathcal{X} \to \mathcal{Z}$ to extract the representations, and a task network $\boldsymbol{T}_\theta : \mathcal{Z} \to \mathcal{Y}$ to perform the downstream task, where ψ and θ denote the network parameters.

6.3.1 Compositionality theory

Finding good latent representations for the task at hand is fundamental in machine learning [28, 29]. Where supervision is available for the latent representations (the ground truth

generative factors) and the downstream task (the ground truth labels), it is natural to train $\boldsymbol{F}_\psi$ and $\boldsymbol{T}_\theta$ with supervised losses as in the Concept Bottleneck Model [234]. However, in practice, it is common that not all the generative factors of the data are known. When there is insufficient supervision for either the latent representations or the downstream task, learning generalisable and interpretable representations is a challenging problem to solve. To tackle this issue, I propose to use compositional equivariance as an inductive bias to learn the latent representations. I later show that with the compositional equivariance, the model can learn representations that are useful for downstream tasks without any supervision, with weak supervision, or with sparse supervision i.e. un-, weakly-, semi-supervised settings.

6.3.1.1 Compositionality

Following [235], I define a compositional representation as satisfying:

$$\boldsymbol{F}_\psi(S \circ \mathbf{X}) = S' \circ \boldsymbol{F}_\psi(\mathbf{X}), \tag{6.1}$$

where S and S' denote the separation operations (e.g. masking operations as discussed in [235]), which are the same operation but operate on different domains. If the representation of the separated generative factor in $\mathbf{X}$ is equivalent to the separated representation of $\mathbf{X}$ using the same separation operation, then the representation $S \circ \boldsymbol{F}_\psi(\mathbf{X})$ is compositional. For example, the separation operation can be masking the image with the masks of objects as in [235]. Typically, designing such separation operations requires knowing the ground truth generative factors.

6.3.1.2 Compositional equivariance

Equivariance is defined as:

$$\boldsymbol{F}_\psi(M_g \circ \mathbf{X}) = M'_g \circ \boldsymbol{F}_\psi(\mathbf{X}), \tag{6.2}$$

where M_g and M'_g denote a set of transformations. Here, $\boldsymbol{F}_\psi(\mathbf{X})$ is equivariant if there exist M_g and M'_g such that the transformations of the input $\mathbf{X}$ that transform the output $\boldsymbol{F}_\psi(\mathbf{X})$ in

the same manner. I then define a **compositionally equivariant** representation as satisfying:

$$\boldsymbol{F}_\psi(M_g \circ S \circ \mathbf{X}) = M'_g \circ S' \circ \boldsymbol{F}_\psi(\mathbf{X}). \tag{6.3}$$

This implies that a representation is compositionally equivariant if it represents a generative factor that is defined by performing the separation operation on $\mathbf{X}$ and there exist transformations that equivariantly affect the factor in the $\mathcal{X}$ space and in the $\mathcal{Z}$ space. In the real world, the generative factors are usually compositionally equivariant. A simple example is that considering car wheels as a generative factor, composing car wheels with other car components (equivalent to performing transformations on the car wheels) can represent different cars, which does not affect the separation of the car wheels from different cars. I claim that when the learnt latent representations satisfy compositional equivariance, the representations approximate well the ground truth generative factors.

6.3.1.3 Compositionally equivariant representations

To learn a compositionally equivariant representation, the key is to find a proper separation operation or its approximation and to design the transformations. Motivated by [57, 236, 237, 238], I assume that it is known that for a group of data samples $\{\mathbf{X_k^1}, \cdots, \mathbf{X_k^{N_k}}\}$, there exists at least one generative factor that is shared across all samples. In this case, comparing $\{\mathbf{X_k^1}, \cdots, \mathbf{X_k^{N_k}}\}$, we can identify the shared factor. If we compose the shared factor with different factors to generate the different data $\{\mathbf{X_k^1}, \cdots, \mathbf{X_k^{N_k}}\}$, this is equivalent to performing transformations on the shared factor. Hence, with the limited information that the data group shares some factors, we can design an objective to train the model to learn compositionally equivariant representations. In particular, for any $i \in \{1, \cdots, N_k\}$ and $h \in \{1, \cdots, N_k\}$, I aim to minimise the compositionally equivariant objective:

$$\mathcal{L}^{i,h} = |\boldsymbol{F}_\psi(\mathbf{X_k^i})_j - \boldsymbol{F}_\psi(\mathbf{X_k^h})_j|_1, \tag{6.4}$$

where j denotes the index of the shared factor. Note that directly minimising Eq. 6.4 requires knowing which factors are shared across the data group, which is a strong assumption, especially for medical data. Hence, it is more feasible to design specific learning objectives or design biases to implicitly minimise Eq. 6.4. In the following, I study several different approaches that implicitly achieve compositional equivariance.

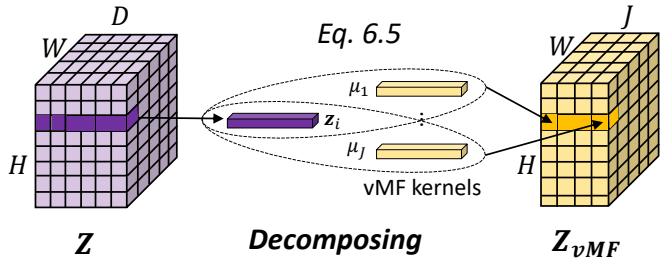


Figure 6.2: The decomposing module. $\mathbf{Z}$ is the features encoded by a feature encoder network. The feature vector $\mathbf{z}_i \in \mathbb{R}^D$ is defined as a vector across channels at position i on the 2D lattice of the feature map. The j^{th} vMF kernel is defined as $\boldsymbol{\mu}_j \in \mathbb{R}^D$. With Eq. 6.5, we can obtain the vMF activations $\mathbf{Z}_{vMF}$. Figure is taken from [17]

6.3.2 Modeling compositional representations

I first model compositional representations with the learnable von-Mises-Fisher (vMF) kernels as shown in Fig. 6.2. In other words, I represent deep features in a compact low dimensional vMF space. I denote the features extracted by $\boldsymbol{F}_\psi$ as $\mathbf{Z} \in \mathbb{R}^{H \times W \times D}$, where H and W are the spatial dimensions and D is the number of channels. The feature vector $\mathbf{z}_i \in \mathbb{R}^D$ is defined as a vector across channels at position i on the 2D lattice of the feature map. I follow Compositional Networks [149] to model $\mathbf{Z}$ with J vMF distributions, where the learnable mean of the j^{th} vMF kernel distribution is defined as $\boldsymbol{\mu}_j \in \mathbb{R}^D$. To make the modelling tractable, the variance σ of all distributions is fixed. In particular, the vMF activation for the j^{th} distribution at each position i can be calculated as:

$$z_{i,j} \equiv p(\mathbf{z}_i|\boldsymbol{\mu}_j) = \frac{e^{\sigma_j \boldsymbol{\mu}_j^T \mathbf{z}_i}}{C(\sigma)}, \text{ s.t. } ||\boldsymbol{\mu}_j|| = 1, \tag{6.5}$$

where the feature vector is normalised i.e. $||\mathbf{z}_i|| = 1$ and $C(\sigma)$ is a constant. The inner product between the feature vectors and the kernels is first calculated, which indicates how much the kernel is activated by each feature vector. After modelling the image features with J vMF distributions according to Eq. 6.5, the tensor of vMF activations $\mathbf{Z}_{vMF} \in \mathbb{R}^{H \times W \times J}$ can be obtained, indicating how much each kernel is activated at each position. Note that it is possible to replace vMF kernels with Gaussian kernels. However, during training, the gradients for updating the kernels and the feature vectors are based on the activations, where vMF activations and the gradients are easier and faster to calculate in implementation (using GPUs) as it simply calculates the inner products and the exponential terms. Moreover, I will introduce the composing

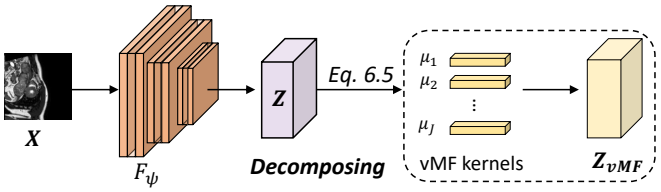


Figure 6.3: Unsupervised compositionally equivariant representation learning model. I train the vMF kernels with Eq. 6.6. $\boldsymbol{F}_\psi$ is the encoding part of a U-Net that is pre-trained to reconstruct the input image.

method later that linearly combines the kernels to represent the original feature vectors, where the original feature space can be well reconstructed with the vMF kernels. I leverage the compositional kernels as compositional representations. However, simply decomposing the features into a compositional latent space does not ensure the assignment of *meaningful* information to each compositional representation i.e. compositional equivariance.

6.3.3 Achieving compositional equivariance

The decomposition process described above allows us to extract compositional representations. However, these representations are not bound to be compositionally equivariant. In other words, the decomposed representations usually do not correspond to the underlying generative factors. In the following, I consider three different settings that can assign generative factors' information to the representations in order to achieve compositional equivariance.

6.3.3.1 Unsupervised setting

I first consider that no supervision information is provided. I use the clustering loss in [12] to enforce the compositional representations to correspond to the centres of any clusters of the input feature vectors (as in Fig. 6.3). The loss $\mathcal{L}_{clu}$ that updates the kernels to be the cluster centres of the feature vectors is defined in [149] as:

$$\mathcal{L}_{clu}(\boldsymbol{\mu}, \mathbf{Z}) = -(HW)^{-1} \sum_i \max_j \boldsymbol{\mu}_j^T \mathbf{z}_i, \tag{6.6}$$

where I only train the kernels and the feature vectors are fixed and produced by the encoding network $\boldsymbol{F}_\psi$. Here, the feature vectors that maximally activate the kernels are first searched.

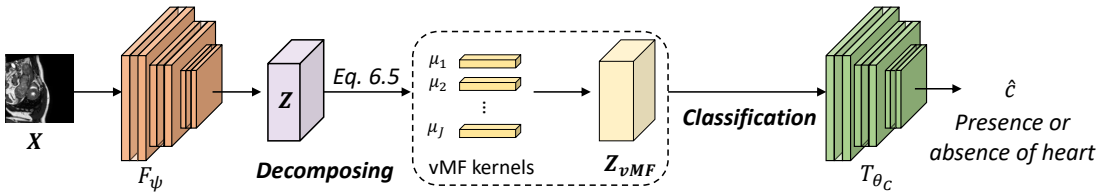


Figure 6.4: Overall model design for weakly supervised compositionally equivariant representation learning. The image is first encoded and then the vMF activations are calculated as the input of the classifier. I use the presence or absence of heart in the image as weak supervision.

Then, the distance between the feature vectors and the kernels is minimised. Note that $\boldsymbol{F}_\psi$ is the encoding part of a U-Net that is pre-trained to reconstruct the input image. If the group of data that shares some factors forms a cluster in latent space, then using the clustering loss will possibly align the kernels with the cluster centres of the data groups. One can expect that the assumption of groups of data forming clusters is not always true in practice. Also, multiple kernels may be aligned to the same cluster centre. It is likely to be that with the clustering loss, the compositional representations can capture part of the information of the factors i.e. achieving a certain level of compositional equivariance.

6.3.3.2 Weakly-supervised setting

Next, I consider using weak supervision describing whether or not a given shared factor is present in each image (e.g. *heart* in cardiac images), as shown in Fig. 6.4. Note that I consider the task of medical image segmentation in this chapter. Hence, to help with downstream tasks, it is important to consider the shared factors that are corresponded to the task. In this case, I can learn compositionally equivariant representations of the heart and potentially use the activations for heart localisation and segmentation. I define the label as c which indicates the presence or absence of the heart in the image. Here, the task network is a binary classifier i.e. $\hat{c} = \boldsymbol{T}_\theta(\mathbf{Z}_{vMF})$. The weak supervision loss to train the model is:

$$\mathcal{L}_{weak}(\hat{c}, c) = |\hat{c} - c|_1. \tag{6.7}$$

I combine this weakly supervised loss with the clustering loss to obtain the overall objective:

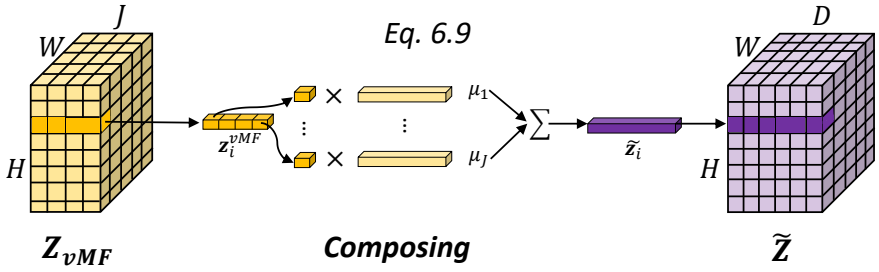


Figure 6.5: The composing module. I construct a new feature space $\widetilde{\mathbf{Z}}$ (with Eq. 6.9) to approximate the encoded features $\mathbf{Z}$, enabling the reconstruction of the input image. Figure is taken from [17] and is reproduced.

$$\underset{\psi,\theta,\boldsymbol{\mu}}{\operatorname{argmin}} \quad \mathcal{L}_{weak}(\hat{c}, c) + \mathcal{L}_{clu}(\boldsymbol{\mu}, \mathbf{Z}). \tag{6.8}$$

After adding weak supervision about the heart, we expect that some of the learned compositional representations will be assigned corresponding information i.e. will be compositionally equivariant representations corresponding to the *heart* factor.

6.3.3.3 Semi-supervised setting with reconstruction

I further consider a semi-supervised setting, by leveraging a reconstruction network $\boldsymbol{R}_\omega$ to train also on data without labels for the downstream segmentation task. As proposed in the conference paper (vMFNet) [12], the model composes the vMF kernels to reconstruct the image with $\boldsymbol{R}_\omega$ by using the vMF activations as the composing operations. Then, the vMF activations that contain spatial information are used to predict the segmentation mask with $\boldsymbol{T}_\theta$. The composing module is shown in Fig. 6.5. The overall model design of the vMFNet is shown in Fig. 6.6. Note that using more unlabeled data in training implicitly constructs more groups of data with the same factors, which enforces the learnt representation to be more compositionally equivariant.

After decomposing the image features with the vMF kernels and the activations, I re-compose to reconstruct the input image. Reconstruction requires that complete information about the input image is captured [56]. In this case, it is possible to observe if the compositional representations have captured information about all the generative factors for the image. However, the vMF activations contain only spatial information, as observed in [149], while style information is compressed as the kernels $\boldsymbol{\mu}_j, j \in \{1 \cdots J\}$, where the compression is not invertible. Consider

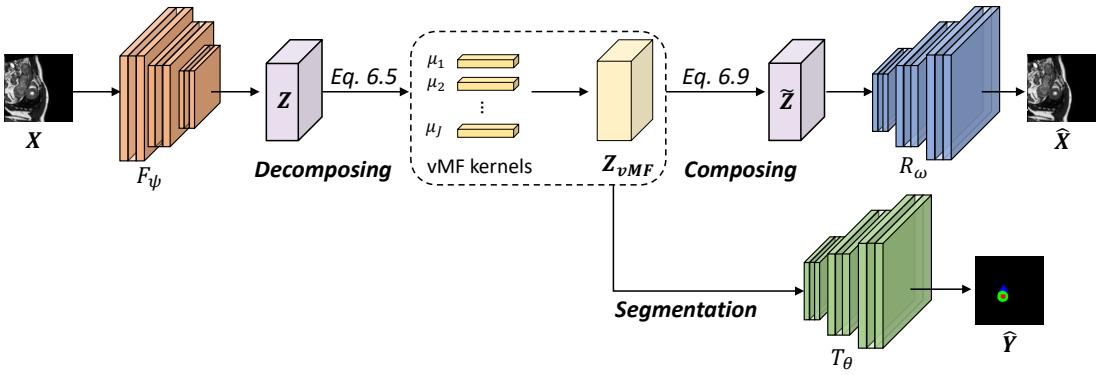


Figure 6.6: Overall model design for semi-supervised compositionally equivariant representation learning for domain-generalised medical image segmentation. The model has been presented in the conference paper [12]. Apart from decomposing and composing modules, the segmentation module is used to predict the segmentation mask by taking the vMF activations as input. Figure is taken from [17] and is reproduced.

that the vMF activation $p(\mathbf{z}_i|\boldsymbol{\mu}_j)$ denotes how much the kernel $\boldsymbol{\mu}_j$ is activated by the feature vector $\mathbf{z}_i$. I construct a new feature space $\widetilde{\mathbf{Z}}$ (as in [12]) with the vMF activations and kernels. Let $\mathbf{z}_i^{vMF} \in \mathbb{R}^J$ be a normalised vector across $\mathbf{Z}_{vMF}$ channels at position i. I devise the new feature vector $\widetilde{\mathbf{z}}_i$ as the combination of the kernels with the normalised vMF activations as the combination coefficients, namely:

$$\widetilde{\mathbf{z}}_i = \sum_{j=1}^{J} \mathbf{z}_{i,j}^{vMF} \boldsymbol{\mu}_j, \text{ where } ||\mathbf{z}_i^{vMF}|| = 1. \tag{6.9}$$

After obtaining $\widetilde{\mathbf{Z}}$ as the approximation of $\mathbf{Z}$, the reconstruction network $\boldsymbol{R}_\omega$ reconstructs the input image with $\widetilde{\mathbf{Z}}$ as the input, i.e. $\hat{\mathbf{X}} = \boldsymbol{R}_\omega(\widetilde{\mathbf{Z}})$. The reconstruction loss is defined as:

$$\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}}) = |\mathbf{X} - \hat{\mathbf{X}}|_1, \tag{6.10}$$

As the vMF activations contain only spatial information of the image that is highly correlated to the segmentation mask, I design a segmentation module, i.e. the task network $\boldsymbol{T}_\theta$, to predict the segmentation mask with the vMF activations as input, i.e. $\hat{\mathbf{Y}} = \boldsymbol{T}_\theta(\mathbf{Z}_{vMF})$. Specifically, the segmentation mask tells what anatomical part the feature vector $\mathbf{z}_i$ corresponds to, which provides further guidance for the model to learn the vMF kernels as the components of the anatomical parts. Then the vMF activations will be further aligned when trained with

multi-domain data and hence perform well on domain generalisation tasks. Overall, the feature vectors of different images corresponding to the same anatomical part will be clustered and activate the same kernels. In other words, the vMF kernels are learnt as the components or patterns of anatomical parts i.e. compositionally equivariant representations. Hence, the vMF activations $\mathbf{Z}_{vMF}$ for the features of different images will be aligned to follow the same distributions (with the same means). In this case, comparing with the content-style disentanglement paradigm [13, 14], the vMF activations can be considered as containing the content information and the vMF kernels as containing the style information.

Overall, the model contains trainable parameters ψ, θ, ω and the kernels $\boldsymbol{\mu}$. The model can be trained end-to-end with the following objective:

$$\underset{\psi,\theta,\omega,\boldsymbol{\mu}}{\operatorname{argmin}} \quad \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y},\hat{\mathbf{Y}}) + \mathcal{L}_{rec}(\mathbf{X},\hat{\mathbf{X}}) + \mathcal{L}_{clu}(\boldsymbol{\mu},\mathbf{Z}), \tag{6.11}$$

where $\lambda_{Dice} = 1$ when the ground truth mask $\mathbf{Y}$ is available, otherwise $\lambda_{Dice} = 0$. $\mathcal{L}_{Dice}$ is the Dice loss as defined in [239].

6.3.3.4 Semi-supervised setting with cross pseudo supervision

An alternative way to take advantage of unlabeled data for the downstream segmentation task is using cross pseudo supervision as proposed in [240]. In particular, I train simultaneously two identical models that are initialised differently, where the pseudo supervision of one model (with networks $\boldsymbol{F}_{\psi}$ and $\boldsymbol{T}_{\theta}$ and kernels $\boldsymbol{\mu}$) is the output of the other model (with networks $\boldsymbol{F}_{\psi'}$ and $\boldsymbol{T}_{\theta'}$ and kernels $\boldsymbol{\mu}'$) with the same input. Such cross pseudo supervision is equivalent to ensembling multiple models to minimise the uncertainty of the prediction. Here, I design the segmentation model by directly using the vMF activations as the input to a segmentation module as shown in Fig. 6.7. The cross pseudo supervision (CPS) loss is defined as:

$$\mathcal{L}_{CPS}(\mathbf{Y}_{pseudo},\hat{\mathbf{Y}}) = \mathcal{L}_{Dice}(\mathbf{Y}_{pseudo},\hat{\mathbf{Y}}), \tag{6.12}$$

where $\mathbf{Y}_{pseudo}$ is the pseudo ground truth segmentation mask and is detached during training (to stop gradients). Overall, the model is trained with the following objective:

$$\begin{aligned}\underset{\psi,\theta,\boldsymbol{\mu},\psi',\theta',\boldsymbol{\mu}'}{\operatorname{argmin}} \quad & \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y},\hat{\mathbf{Y}}) + \mathcal{L}_{clu}(\boldsymbol{\mu},\mathbf{Z}')+ \\ & \lambda_{Dice}\mathcal{L}_{Dice}(\mathbf{Y},\hat{\mathbf{Y}}') + \mathcal{L}_{clu}(\boldsymbol{\mu}',\mathbf{Z}')+ \\ & \lambda_{CPS}\mathcal{L}_{CPS}(\hat{\mathbf{Y}},\hat{\mathbf{Y}}') + \lambda_{CPS}\mathcal{L}_{CPS}(\hat{\mathbf{Y}}',\hat{\mathbf{Y}}),\end{aligned} \tag{6.13}$$

where $\lambda_{Dice} = 1$ when the ground truth mask $\mathbf{Y}$ is available, otherwise $\lambda_{Dice} = 0$. I set λ_{CPS} as 0.1. The model is termed vMFPseudo.

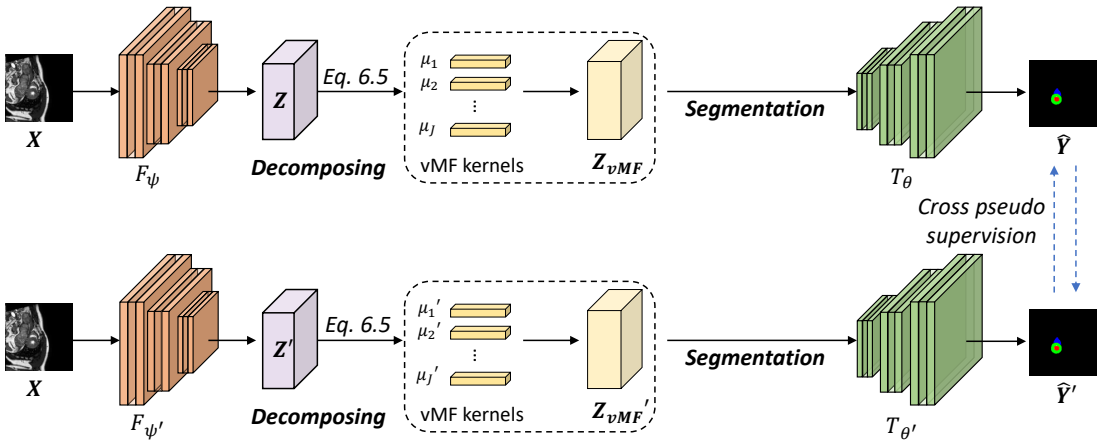


Figure 6.7: Overall model design for semi-supervised compositionally equivariant representation learning with cross pseudo supervision for domain-generalised medical image segmentation. I simultaneously train two models and use the prediction of one model as the pseudo supervision for the other model. The segmentation module is used to predict the segmentation mask by taking the vMF activations as input.

6.4 Experiments

For all the experiments, I adopt **multi-centre, multi-vendor & multi-disease cardiac image segmentation (M&Ms) dataset [5]** and **spinal cord gray matter segmentation (SCGM) dataset [7]**.

6.4.1 Implementation details

All models are trained using the Adam optimiser [180] with a learning rate of $1 \times e^{-4}$ for 50K iterations using a batch size of 4 for the semi-supervised settings. Images are cropped to 288×288 for M&Ms and 144×144 for SCGM.

$\boldsymbol{F}_\psi$ contains all the downsampling and part of the upsampling layers of a 2D U-Net [218] to extract features $\mathbf{Z}$, where the last upsampling and output layers are dropped and the skip connections are reserved between the downsampling and upsampling layers. The last upsampling layers are dropped. Note that $\boldsymbol{F}_\psi$ can be replaced by other encoders such as a ResNet [241] and the feature vectors can be extracted from any layer of the encoder where performance may vary for different layers. For all settings, I pre-train the U-Net for 50 epochs with unlabeled data from the source domains. For the weakly supervised setting, the classifier $\boldsymbol{T}_\theta$ has 5 CONV-BN-LeakyReLU layers (kernel size 4, stride size 2 and padding size 1) and two fully-connected layers that down-sample the features to 16 dimensions and 1 dimension (for output). For the semi-supervised settings, $\boldsymbol{T}_\theta$ and $\boldsymbol{R}_\omega$ have similar structures, where a double CONV layer (kernel size 3, stride size 1 and padding size 1) in U-Net with batch normalisation and ReLU is first used to process the features. Then a transposed convolutional layer is used to upsample the features followed by a double CONV layer with batch normalisation and ReLU. Finally, an output convolutional layer with 1×1 kernels is used. For $\boldsymbol{T}_\theta$, the output of the last layer is processed with a sigmoid operation.

I follow [149] to set the variance of the vMF distributions to 30. The number of kernels is set to 12, as it was found empirically in early experiments that this number performed the best. For different medical datasets, the best number of kernels may be slightly different. All models are implemented in PyTorch [54] and are trained using an NVIDIA 2080 Ti GPU. In semi-supervised settings, I use specific percentages of the subjects as labelled data and the rest as unlabeled data. I train the models with 3 source domains and treat the 4^{th} domain as the target one. I use Dice (expressed as %) [195] and Hausdorff Distance (HD) [228] as the evaluation metrics.

6.4.2 How to evaluate compositional equivariance?

The generative factors are considered to be generalisable and interpretable. I hence consider how interpretable the activations of the compositionally equivariant representations are and how generalisable the representations are. For interpretability, I follow [14] to consider how much each vMF activation channel is meaningful (carries information that is relevant to specific anatomy) and how homologous each channel is. For generalisation ability, I consider the performance of the model on the task of semi-supervised domain generalisation as in [12].

6.4.3 Unsupervised setting

I train the model as shown in Fig. 6.3 with Eq. 6.6 for 200 epochs with all the labelled data of the M&Ms dataset. I show the qualitative results for the unsupervised setting in Fig. 6.8. With only the clustering loss, some channels are already meaningful i.e. corresponding to specific anatomy. For example, channel 1 (red box) contains information on the left ventricle (LV) and right ventricle (RV) of the heart. Part of channel 2 is relevant to the lungs. Channel 3 corresponds to the background.

6.4.4 Weakly-supervised setting

For the weakly supervised setting, I train the model as shown in Fig. 6.4 with Eq. 6.8 for 200 epochs with all the labelled data of M&Ms dataset. The qualitative results are shown in Fig. 6.9. It is clearly shown that a stronger compositional equivariance is achieved compared to the unsupervised setting. Channels 1 and 2 (red box) are more related to the heart. Channel 3 (yellow box) shows the shape of the lungs even without any supervision on the task. Channel 4 contains mostly the background. Overall, the activations of the compositional representations are more interpretable and each channel is more homologous i.e. more compositionally equivariant. Interestingly, for both unsupervised and weakly supervised settings, I observe that one compositional representation represents the lungs even though no information about the lungs is provided. This means that the learnt representations are ready to be used for lung localisation/segmentation when there is a small amount of relevant labelled data available i.e. robust to the task of lung localisation/segmentation.

6.4.5 Semi-supervised setting with reconstruction

For the semi-supervised settings, I test the methods on semi-supervised domain generalisation problems.

6.4.5.1 Baseline models

For a fair comparison, I compare all models with the same backbone feature extractor, i.e. U-Net [218], without any pre-training on other datasets. The baseline models have been described in Chapter 5. DGNet [16] is the semi-supervised gradient-based meta-learning approach intro-

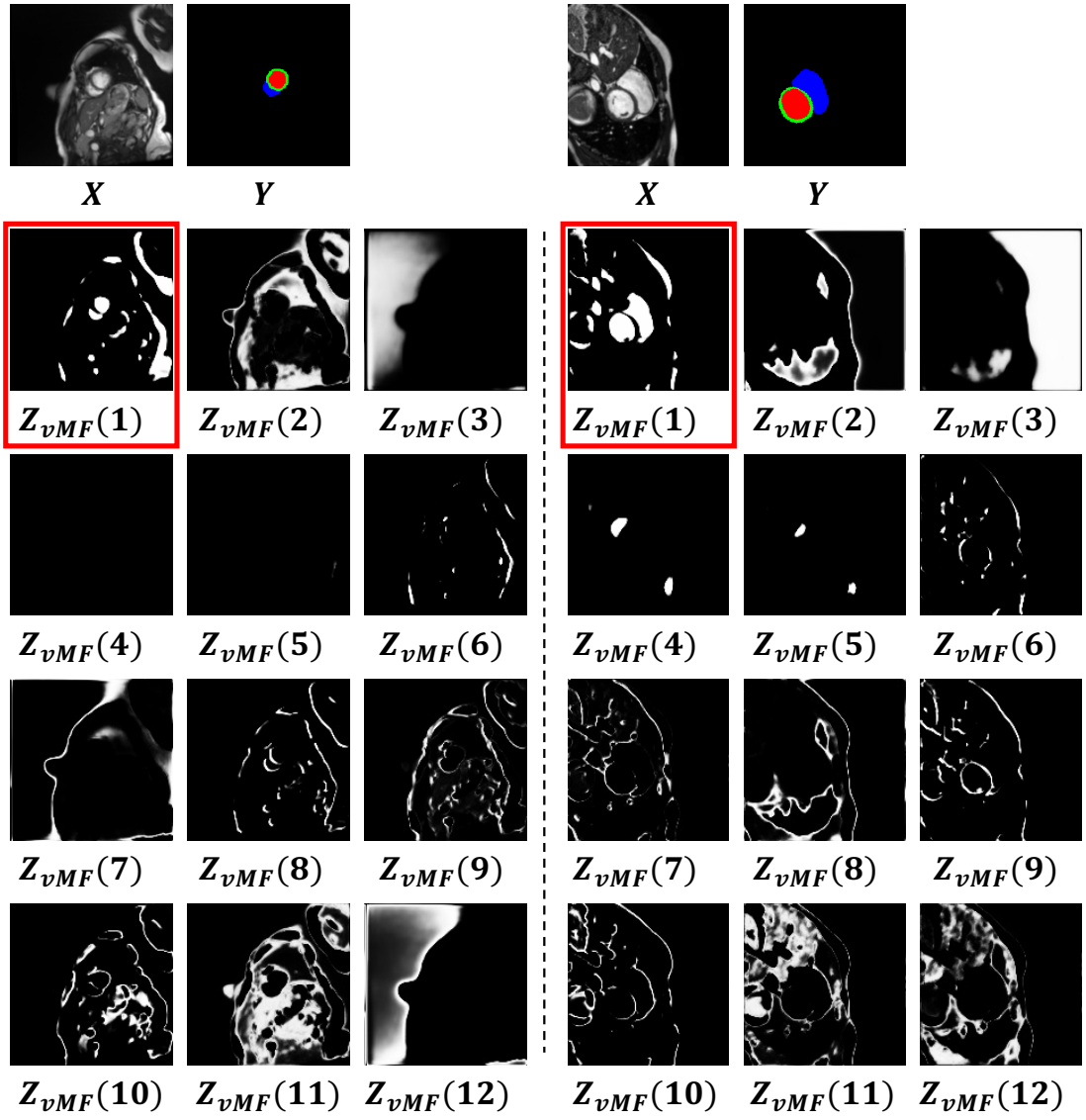


Figure 6.8: Visualisation of images, ground truth segmentation masks, and 12 vMF activation channels for 2 example images using **the unsupervised setting** from M&Ms dataset. The channels are manually ordered. The red box highlights the activation of the kernel (partially) corresponding to the heart.

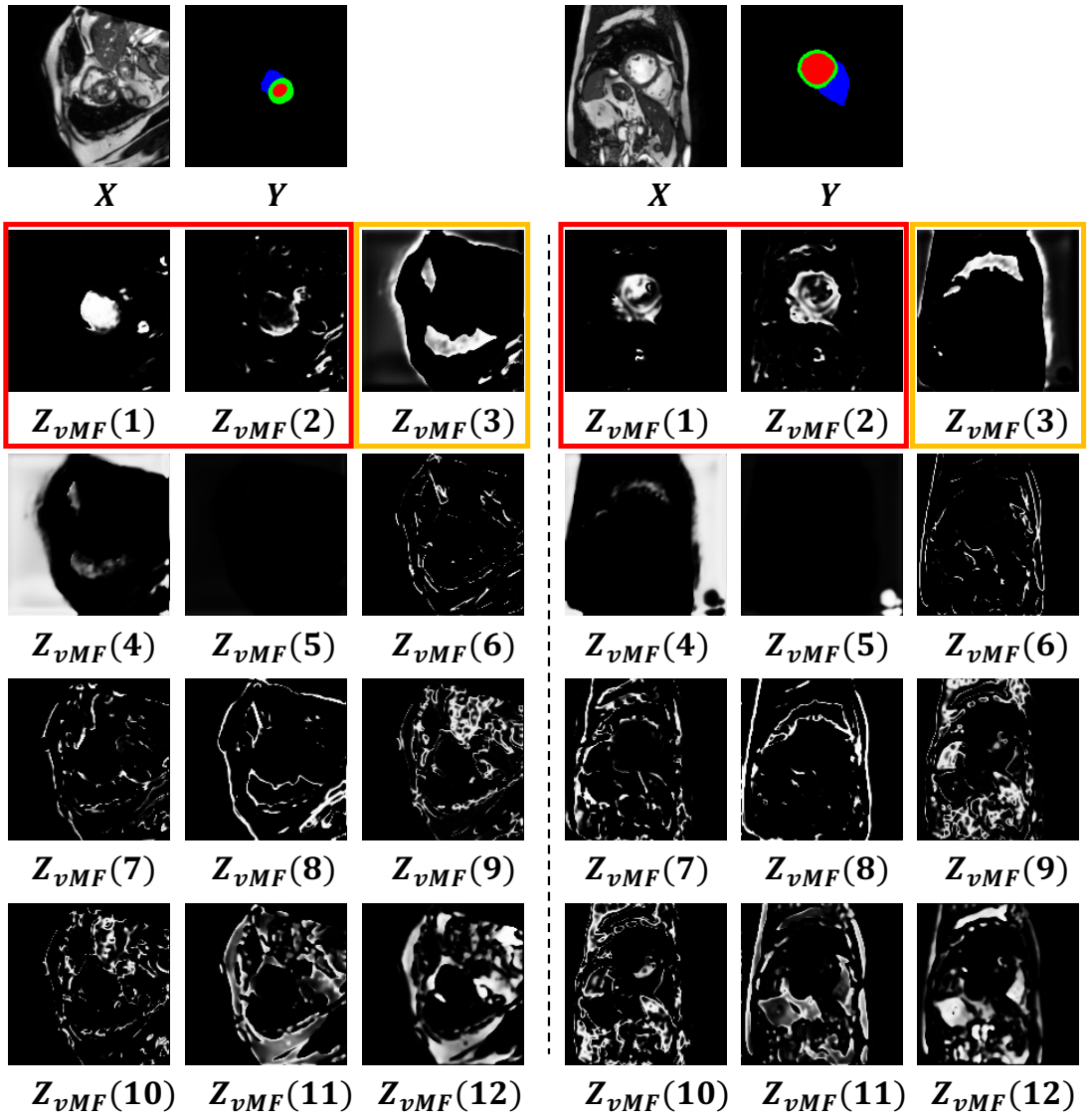


Figure 6.9: Visualisation of images, ground truth segmentation masks, and 12 vMF activation channels for 2 examples of **the weakly supervised setting** from M&Ms dataset. The channels are manually ordered. The red box highlights the activation of the kernel (partially) corresponding to the heart. The yellow box relates to the channel that contains information about the lungs.

duced in Chapter 5.

6.4.5.2 Generalisation

Table 6.1 reports the average results over four leave-one-out experiments that treat each domain in turn as the target domain; more detailed results can be found in Tables 6.2 – 6.5. I highlight that the proposed vMFNet is **14 times faster to train** compared to the previous SOTA DGNet. Training vMFNet for one epoch takes 7 minutes, while DGNet needs 100 minutes for the M&Ms dataset due to the need to construct new computational graphs for the meta-test step in every iteration.

With limited annotations, vMFNet achieves 7.7% and 3.0% improvements (in Dice) for 2% and 5% cases compared to the previous SOTA DGNet on M&Ms dataset. For the 100% case, vMFNet and DGNet have similar performance of around 86% Dice and 14 HD. Overall, vMFNet has consistently better performance for almost all scenarios on the M&Ms dataset. Similar improvements are observed for the SCGM dataset.

6.4.5.3 Interpretability

Overall, the segmentation prediction can be interpreted as the activation of corresponding compositional representations (kernels) at each position, where false predictions occur when the wrong representations are activated i.e. the wrong vMF activations are used to predict the mask. I show example images, reconstructions, predicted segmentation masks, and the 12 vMF activations channels in Fig. 6.11. As shown, channels 1 and 2 (red box) are mostly activated by LV feature vectors and channels 3 (blue box) and 4 (green box) are mostly for RV and myocardium (MYO) feature vectors. Interestingly, channel 2 is mostly activated by papillary muscles in the left ventricle even though no supervision about the papillary muscles is provided during training. This supports that the model learns the kernels as the compositionally equivariant representations (patterns of papillary muscles, LV, RV and MYO) of the heart. Although part of channel 10 corresponds to the lungs, the other channels (e.g. channels 8-12) contain mixed (not interpretable and homologous) information about the image as the representations have to contain complete information about the image.

Percent	metrics	nnUNet	SDNet+Aug.	LDDG	SAML	DGNet	vMFPseudo	vMFNet
M&Ms 2%	Dice (↑)	$65.94_{8.3}$	$68.28_{8.6}$	$63.16_{5.4}$	$64.57_{8.5}$	$72.85_{4.3}$	$77.97_{4.7}$	$\mathbf{78.43_{3.6}}$
	HD (↓)	$20.96_{4.0}$	$20.17_{3.3}$	$22.02_{3.5}$	$21.22_{4.1}$	$19.32_{2.8}$	$16.61_{1.8}$	$\mathbf{16.56_{1.7}}$
M&Ms 5%	Dice (↑)	$76.09_{6.3}$	$77.47_{3.9}$	$71.29_{3.6}$	$74.88_{4.6}$	$79.75_{4.4}$	$\mathbf{82.55_{2.6}}$	$82.12_{3.1}$
	HD (↓)	$18.22_{3.0}$	$18.62_{3.1}$	$19.21_{3.0}$	$18.49_{2.9}$	$17.98_{3.2}$	$\mathbf{15.10_{1.5}}$	$15.30_{1.8}$
M&Ms 100%	Dice (↑)	$84.87_{2.5}$	$84.29_{1.6}$	$85.38_{1.6}$	$83.49_{1.3}$	$\mathbf{86.03_{1.7}}$	$85.49_{1.6}$	$85.92_{2.0}$
	HD (↓)	$14.80_{1.9}$	$15.06_{1.6}$	$14.88_{1.7}$	$15.52_{1.5}$	$14.53_{1.8}$	$\mathbf{13.99_{1.1}}$	$14.05_{1.3}$
SCGM 20%	Dice (↑)	$64.85_{5.2}$	76.73_{11}	63.31_{17}	73.50_{12}	79.58_{11}	75.58_{11}	$\mathbf{81.11_{8.8}}$
	HD (↓)	$3.49_{0.49}$	$2.07_{0.36}$	$2.38_{0.39}$	$2.11_{0.37}$	$1.97_{0.30}$	$2.17_{0.36}$	$\mathbf{1.96_{0.31}}$
SCGM 100%	Dice (↑)	$71.51_{5.4}$	81.37_{11}	79.29_{13}	80.95_{13}	82.25_{11}	$\mathbf{85.01_{5.8}}$	$84.03_{8.0}$
	HD (↓)	$3.53_{0.45}$	$1.93_{0.36}$	$2.11_{0.41}$	$1.95_{0.38}$	$1.92_{0.31}$	$1.89_{0.25}$	$\mathbf{1.84_{0.31}}$

Table 6.1: Average Dice (%) and Hausdorff Distance (HD) results and the standard deviations on M&Ms and SCGM datasets. For semi-supervised approaches, the training data contain all unlabeled data and different percentages of labelled data from source domains. The rest are trained with different percentages of labelled data only. Results of baseline models are taken from [12]. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	DGNet	vMFPseudo	vMFNet
2%	B,C,D	A	52.87_{19}	54.48_{18}	59.47_{12}	56.31_{13}	66.01_{12}	70.12_{16}	$\mathbf{73.13_{9.6}}$
	A,C,D	B	64.63_{17}	67.81_{14}	56.16_{14}	56.32_{15}	72.72_{10}	$\mathbf{78.77_{10}}$	$77.01_{7.9}$
	A,B,D	C	72.97_{14}	76.46_{12}	68.21_{11}	$75.70_{8.7}$	77.54_{10}	$\mathbf{81.75_{8.6}}$	$81.57_{8.1}$
	A,B,C	D	73.27_{11}	74.35_{11}	68.56_{10}	$69.94_{9.8}$	$75.14_{8.4}$	$81.23_{7.0}$	$\mathbf{82.02_{6.5}}$
5%	B,C,D	A	65.30_{17}	71.21_{13}	$66.22_{9.1}$	67.11_{10}	72.40_{12}	$\mathbf{78.06_{8.8}}$	77.06_{10}
	A,C,D	B	79.73_{10}	77.31_{10}	$69.49_{8.3}$	$76.35_{7.9}$	$80.30_{9.1}$	$\mathbf{83.49_{7.1}}$	$82.29_{7.8}$
	A,B,D	C	78.06_{11}	$81.40_{8.0}$	$73.40_{9.8}$	$77.43_{8.3}$	$82.51_{6.6}$	$83.71_{7.3}$	$\mathbf{84.01_{7.3}}$
	A,B,C	D	$81.25_{8.3}$	$79.95_{7.8}$	$75.66_{8.5}$	$78.64_{5.8}$	$83.77_{5.1}$	$84.93_{6.1}$	$\mathbf{85.13_{6.1}}$
100%	B,C,D	A	80.84_{11}	$81.50_{7.7}$	$82.62_{6.3}$	$81.33_{7.2}$	$\mathbf{83.21_{7.4}}$	$82.72_{7.1}$	$82.67_{7.2}$
	A,C,D	B	$\mathbf{86.76_{5.8}}$	$85.04_{6.1}$	$85.68_{5.7}$	$84.15_{5.9}$	$86.53_{5.3}$	$86.56_{4.9}$	$85.95_{5.6}$
	A,B,D	C	$84.92_{7.1}$	$85.64_{6.5}$	$86.49_{6.3}$	$84.52_{6.2}$	$87.22_{6.1}$	$85.86_{7.5}$	$\mathbf{87.80_{4.4}}$
	A,B,C	D	$86.94_{5.9}$	$84.96_{5.2}$	$86.73_{6.1}$	$83.96_{5.9}$	$87.16_{4.9}$	$86.81_{4.5}$	$\mathbf{87.26_{4.7}}$

Table 6.2: Dice (%) results and the standard deviations on M&Ms dataset. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	DGNet	vMFPseudo	vMFNet
2%	B,C,D	A	$26.48_{7.5}$	$24.69_{7.0}$	$25.56_{5.9}$	$25.57_{5.7}$	$23.55_{6.5}$	$19.51_{6.2}$	$\mathbf{19.14_{4.8}}$
	A,C,D	B	$23.11_{6.8}$	$21.84_{6.2}$	$25.44_{5.2}$	$24.91_{5.5}$	$19.95_{6.3}$	$\mathbf{16.84_{5.3}}$	$17.01_{3.7}$
	A,B,D	C	$16.75_{4.6}$	$16.57_{4.2}$	$18.98_{3.9}$	$16.46_{3.5}$	$16.29_{4.0}$	$\mathbf{15.06_{3.7}}$	$15.30_{3.5}$
	A,B,C	D	$17.51_{4.9}$	$17.57_{4.1}$	$18.08_{3.8}$	$17.94_{3.8}$	$17.48_{4.7}$	$15.04_{3.2}$	$\mathbf{14.80_{3.0}}$
5%	B,C,D	A	$23.04_{6.7}$	$22.84_{6.3}$	$23.35_{5.7}$	$23.10_{5.9}$	$22.55_{6.6}$	$\mathbf{17.54_{4.9}}$	$18.19_{4.9}$
	A,C,D	B	$18.18_{4.7}$	$20.26_{5.5}$	$20.56_{4.7}$	$18.97_{4.9}$	$19.37_{6.4}$	$\mathbf{14.86_{4.2}}$	$15.24_{3.2}$
	A,B,D	C	$16.44_{4.2}$	$16.22_{3.9}$	$17.14_{3.3}$	$16.29_{3.2}$	$15.77_{3.8}$	$14.35_{3.3}$	$\mathbf{14.17_{3.3}}$
	A,B,C	D	$15.24_{4.2}$	$15.15_{3.3}$	$15.80_{3.2}$	$15.58_{3.2}$	$14.24_{2.8}$	$13.64_{2.8}$	$\mathbf{13.61_{2.8}}$
100%	B,C,D	A	$17.86_{5.5}$	$17.39_{4.5}$	$17.48_{4.1}$	$17.70_{4.2}$	$17.28_{3.9}$	$\mathbf{15.82_{3.9}}$	$15.99_{3.5}$
	A,C,D	B	$14.82_{3.4}$	$15.55_{3.7}$	$15.42_{3.4}$	$16.05_{3.7}$	$14.99_{3.6}$	$\mathbf{13.94_{3.2}}$	$14.58_{3.2}$
	A,B,D	C	$13.72_{3.3}$	$13.67_{3.0}$	$13.52_{2.8}$	$14.21_{3.3}$	$13.11_{2.8}$	$13.12_{3.1}$	$\mathbf{12.70_{2.8}}$
	A,B,C	D	$12.81_{3.4}$	$13.64_{2.9}$	$13.11_{3.0}$	$14.12_{2.8}$	$\mathbf{12.72_{2.6}}$	$13.07_{2.5}$	$12.94_{2.5}$

Table 6.3: Hausdorff Distance results and the standard deviations on M&Ms dataset. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	DGNet	vMFPseudo	vMFNet
20%	2,3,4	1	59.07_{21}	83.07_{16}	$77.71_{9.1}$	78.71_{25}	$87.45_{6.3}$	$87.64_{8.8}$	$\mathbf{88.08_{6.9}}$
	1,3,4	2	69.94_{12}	$80.01_{5.2}$	44.08_{12}	75.58_{12}	$81.05_{5.2}$	63.50_{16}	$\mathbf{81.21_{4.2}}$
	1,2,4	3	$60.25_{7.2}$	58.57_{10}	$48.04_{5.5}$	$54.36_{7.6}$	$61.85_{7.3}$	$64.84_{9.3}$	$\mathbf{66.74_{4.9}}$
	1,2,3	4	$70.13_{4.3}$	$85.27_{2.2}$	$83.42_{2.7}$	$85.36_{2.8}$	$87.96_{2.1}$	$86.35_{2.8}$	$\mathbf{88.39_{2.4}}$
100%	2,3,4	1	$75.27_{8.3}$	$90.25_{4.5}$	$88.21_{4.9}$	$90.22_{5.6}$	$90.01_{4.9}$	$89.78_{4.7}$	$\mathbf{90.96_{4.7}}$
	1,3,4	2	$76.32_{2.9}$	$84.13_{4.2}$	$83.76_{3.1}$	$\mathbf{86.65_{3.5}}$	$85.48_{2.3}$	$83.39_{4.8}$	$84.89_{3.2}$
	1,2,4	3	$62.59_{6.9}$	62.18_{10}	$56.11_{9.3}$	$58.27_{9.4}$	$64.23_{9.7}$	$\mathbf{76.27_{3.7}}$	$70.71_{9.2}$
	1,2,3	4	$71.87_{2.5}$	$88.93_{1.9}$	$89.08_{2.7}$	$88.66_{2.6}$	$89.26_{2.5}$	$\mathbf{90.60_{2.0}}$	$89.57_{3.1}$

Table 6.4: Dice (%) results and the standard deviations on SCGM dataset. Bold numbers denote the best performance.

Source		Target	nnUNet	SDNet+Aug.	LDDG	SAML	DGNet	vMFPseudo	vMFNet
20%	2,3,4	1	$3.09_{0.25}$	$1.52_{0.33}$	$1.75_{0.26}$	$1.53_{0.38}$	$1.50_{0.30}$	$1.55_{0.34}$	$\mathbf{1.47_{0.33}}$
	1,3,4	2	$3.16_{0.09}$	$1.97_{0.16}$	$2.73_{0.33}$	$2.07_{0.35}$	$\mathbf{1.91_{0.16}}$	$2.40_{0.39}$	$1.92_{0.14}$
	1,2,4	3	$3.38_{0.27}$	$2.45_{0.27}$	$2.67_{0.25}$	$2.52_{0.24}$	$\mathbf{2.23_{0.23}}$	$2.43_{0.31}$	$2.25_{0.16}$
	1,2,3	4	$4.31_{0.14}$	$2.34_{0.21}$	$2.37_{0.14}$	$2.30_{0.18}$	$2.22_{0.13}$	$2.30_{0.19}$	$\mathbf{2.18_{0.14}}$
100%	2,3,4	1	$3.26_{0.21}$	$1.37_{0.25}$	$1.50_{0.23}$	$1.43_{0.36}$	$1.43_{0.29}$	$1.49_{0.32}$	$\mathbf{1.35_{0.25}}$
	1,3,4	2	$3.19_{0.09}$	$1.88_{0.16}$	$2.19_{0.19}$	$\mathbf{1.80_{0.19}}$	$1.81_{0.15}$	$1.88_{0.17}$	$\mathbf{1.80_{0.19}}$
	1,2,4	3	$3.37_{0.27}$	$2.34_{0.24}$	$2.64_{0.28}$	$2.43_{0.33}$	$2.23_{0.32}$	$\mathbf{2.13_{0.21}}$	$\mathbf{2.13_{0.30}}$
	1,2,3	4	$4.30_{0.15}$	$2.13_{0.17}$	$2.12_{0.15}$	$2.15_{0.15}$	$2.11_{0.13}$	$\mathbf{2.06_{0.17}}$	$2.07_{0.18}$

Table 6.5: Hausdorff Distance results and the standard deviations on SCGM dataset. Bold numbers denote the best performance.

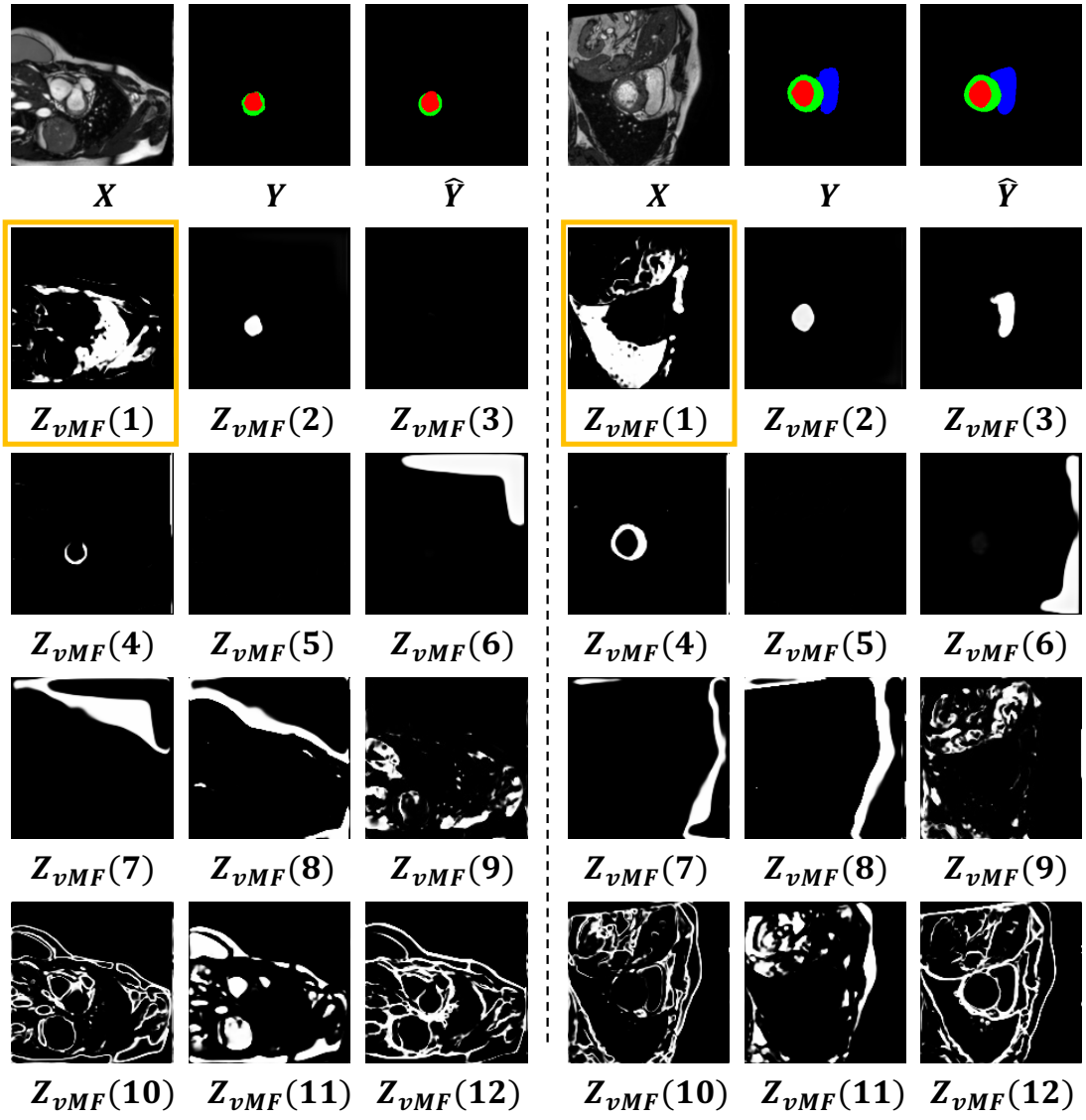


Figure 6.10: Visualisation of images, ground truth segmentation masks, predicted segmentation masks and 12 vMF activation channels for 2 examples of **vMFPseudo** from M&Ms dataset. The channels are manually ordered. The yellow box highlights the channel that contains information about the lungs.

Percent	B,C,D→ A		A,C,D→ B		A,B,D → C		A,B,C→ D		Average		Improvement	
2%	77.11	18.98	80.23	16.82	83.06	14.85	84.29	14.64	81.17	16.32	3.5%	0.14%
5%	78.97	17.97	83.83	15.18	84.13	14.29	86.04	13.45	83.24	15.22	1.4%	0.05%
100%	83.98	15.76	85.75	14.28	88.57	12.39	88.64	12.66	86.74	13.77	0.95%	0.2%

Table 6.6: Dice (%) | Hausdorff Distance results on M&Ms dataset with test-time training. Improvements are the comparison between the average results of with or without test-time training.

6.4.5.4 Which losses help more?

I ablate two key losses of vMFNet in the 2% of M&Ms setting. Note that both losses do not require the ground truth masks. Removing $\mathcal{L}_{rec}$ results in 74.83% Dice and 18.57 HD, whereas removing $\mathcal{L}_{vMF}$ gives 75.45% Dice and 17.53 HD. Removing both gives 74.70% Dice and 18.25 HD. Compared with 78.43% Dice and 16.56 HD, training with both losses gives better generalisation results when the model is trained to learn better kernels and with unlabeled data. When removing the two losses, the model can still perform adequately compared to the baselines due to the decomposing mechanism.

6.4.5.5 Alignment analysis

To show that the vMF likelihoods from different source domains are aligned, for M&Ms 100% cases, I first mask out the non-heart part of the images, features and vMF likelihoods. Then, I train classifiers to predict which domain the input is from with the masked images $\mathbf{X}$ or masked features $\mathbf{Z}$ or masked vMF likelihoods $\mathbf{Z}_{vMF}$ as input. The average cross-entropy errors are 0.718, 0.701 and 0.756, which means that it is harder to tell the domain class with the heart part of $\mathbf{Z}_{vMF}$, i.e. the vMF likelihoods for the downstream task are better aligned compared to the features $\mathbf{Z}$ from different source domains.

6.4.5.6 Test-time domain generalisation

As discussed in Section 6.4.5.3, poor segmentation predictions are usually caused by the wrong kernels being activated. This results in the wrong vMF likelihoods being used to predict masks. The reconstruction quality is also affected by wrong vMF likelihoods. In fact, the average reconstruction error is approximately 0.007 on the training set and 0.011 on the test set. Inspired

by [242, 243] I perform test-time training (TTT) to better reconstruct by fine-tuning the reconstruction loss $\mathcal{L}_{rec}(\mathbf{X}, \hat{\mathbf{X}})$ to update $\boldsymbol{F}_\psi$ and $\boldsymbol{R}_\omega$ with the kernels and $\boldsymbol{T}_\theta$ fixed. This should in turn produce better vMF likelihoods. For images of each subject at test time, I fine-tune the reconstruction loss for 15 iterations (saving the model at each iteration) with a small learning rate of $1 \times e^{-6}$. Out of the 15 models, I choose the one with minimum reconstruction error to predict the segmentation masks for each subject. The detailed results of TTT for M&Ms are included in Table 6.6. For M&Ms 2%, 5% and 100% cases, TTT gives around 3.5%, 1.4% and 1% improvements in Dice compared to results (without TTT) in Table 6.1.

6.4.6 Semi-supervised setting with pseudo supervision

6.4.6.1 Generalisation

The results of vMFPseudo can be found in Table 6.1, Table 6.2, Table 6.3, Table 6.4 and Table 6.5. It is notable that vMFPseudo has a similar advantage in the computational load and training speed as vMFNet compared to DGNet. Training vMFPseudo for one epoch takes around 14 minutes, while DGNet needs 100 minutes for the M&Ms dataset.

Similar to the improvement of vMFNet over the previous SOTA DGNet, vMFPseudo achieves 7.0% and 3.5% improvements (in Dice) for 2% and 5% cases on the M&Ms dataset. For the 100% case, vMFNet is slightly worse than DGNet and vMFNet, which is around 85.5% Dice and 14 HD. Overall, vMFPseudo consistently performs better for most of the cases compared to the baseline methods for the M&Ms dataset and SCGM dataset. Compared to vMFNet, I observe that for some cases (e.g. 5% B,C,D→A on the M&Ms dataset and 100% 1,2,4→3 on the SCGM dataset), vMFPseudo has clearly better performance. Note that the domain difference between the source domains and the target domain is relatively larger than that in other cases. Hence, the model may produce highly uncertain results for some images in the target domain. In these cases, the cross pseudo supervision loss may help more in mitigating the uncertainty, which produces better results.

6.4.6.2 Interpretability

Overall, I observe more interpretable results with vMFPseudo. First of all, the lungs in the images are more clearly shown in channel 1 (yellow box), which means better robustness regarding generalising to other tasks. Channels 1-3 correspond to LV, RV and MYO. As no

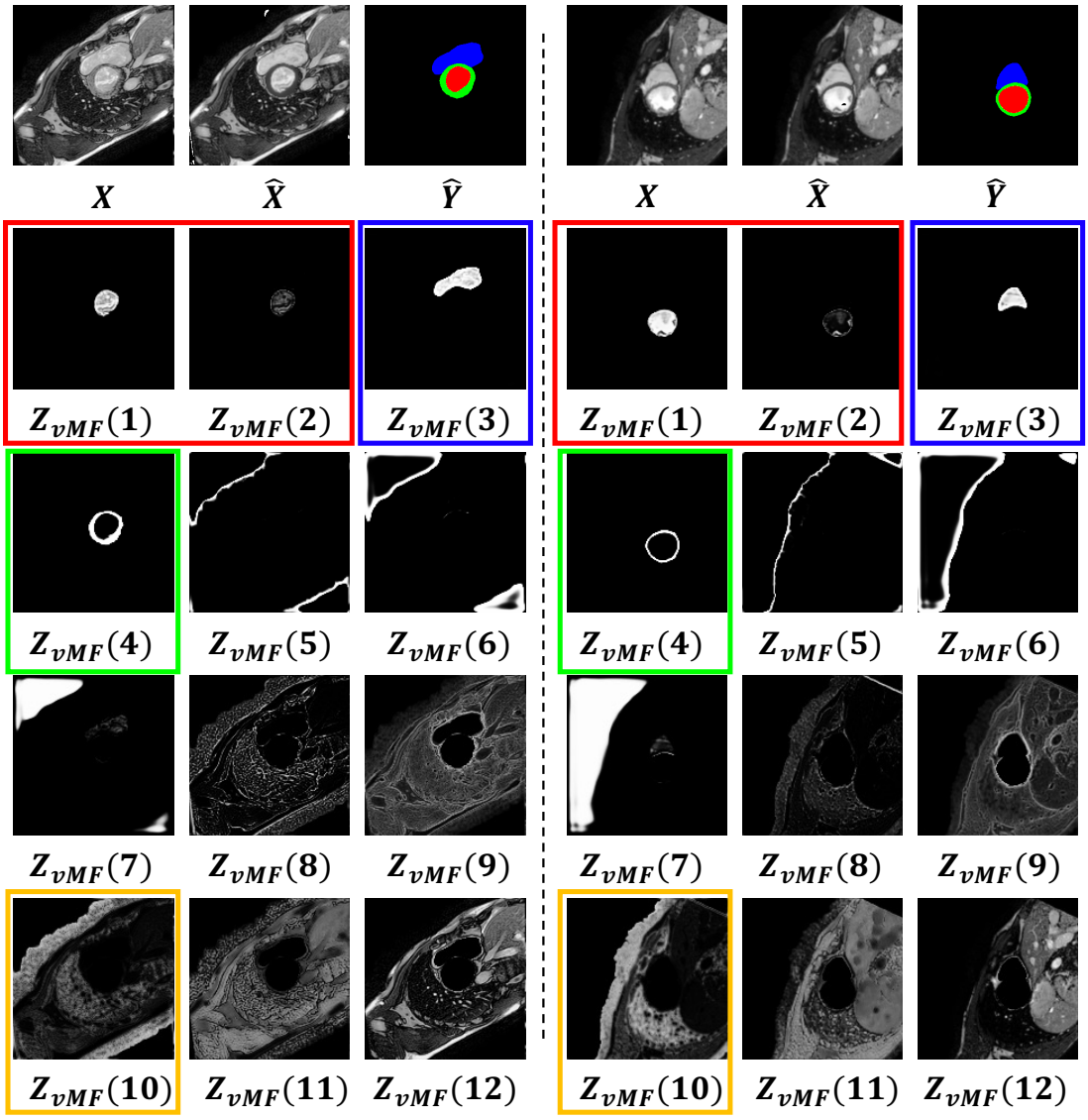


Figure 6.11: Visualisation of images, reconstructions, predicted segmentation masks and 12 vMF activation channels for 2 examples of **vMFNet** from M&Ms dataset. The channels are manually ordered. The red box, blue box and green box highlight the activation of the kernels corresponding to the left ventricle, right ventricle and myocardium. The yellow box relates to the channel that contains information about the lungs.

reconstruction is needed for vMFPseudo, we can see that the other channels are more homologous. For example, channel 10 may relate to the contours of the images.

6.5 Summary

In this Chapter, I have presented that using compositional equivariance as an inductive bias helps to learn generalisable and interpretable compositional representations. In particular, I used different learning biases in different settings to constrain the representations to be compositionally equivariant. For the unsupervised setting and weakly supervised setting, I observed that the representations achieve a certain level of compositional equivalence, which is partially interpretable. For the semi-supervised settings, I qualitatively showed that some of the representations are well interpretable when little supervision is given. Quantitatively, vMFNet and vMFPseudo, the models built based on decomposing the compositional representations with different design biases and learning biases, achieved the best generalisation performance compared to several strong baselines. Overall, as I discussed in Section 6.3 and demonstrated with the results, different learning settings and biases allow the model to learn the representations that are compositionally equivariant at different levels. I conclude that strong prior knowledge (e.g. weak supervision "*heart or not*") or some supervision significantly improves the ability to achieve compositional equivariance. Taking advantage of the unlabeled data also plays a key role to learn compositionally equivariant representations as it implicitly constructs more groups of data that have shared factors.

Chapter 7
Summary, Limitations and Future Directions

In the final chapter, I summarise the thesis contributions and discusses the significance of our work in Section 7.1. I present the limitations and opportunities in Section 7.2. The future directions inspired by the conducted work are discussed in Section 7.3.

7.1 Summary

This thesis focuses on learning representations for more generalisable solutions to medical image analysis. I first conducted a comprehensive survey on disentangled representation learning and proposed two metrics to measure content-style disentanglement to systematically study the effect of different biases. Then, two disentanglement-based generalisation solutions are proposed to handle the domain shifts between multi-domain data that are collected from different clinical centres or hospitals. Finally, I considered compositionality as a prior to learn more generalisable and interpretable representations.

In Chapter 4, I evaluated the disentanglement between image content and style through experimenting on 3 state-of-the-art models, and showcased how design and learning biases affect disentanglement and by extension task performance. The findings suggest that whilst content-style disentanglement enables the implementation of certain equivariant tasks, partially (dis)entangled can lead to better performance than fully disentangled ones. Additionally, the analysis suggests that strict design constraints on the content space lead to increased interpretability, which could be exploited in post-hoc tasks. Using the findings and the presented metrics will enable the design of better models that achieve the degree of disentanglement that maximises performance, rather than blindly pursuing very high (or low) disentanglement.

In Chapter 5, two data augmentation methods were first proposed to address the domain adaptation and generalisation problems in the field of cardiac image segmentation. In particular, a geometry-related augmentation method was introduced, which aims to remove the scale and

resolution bias of the original data. Further, the second proposed augmentation method aims to bridge the gap between populations and data captured by different scanners. To achieve this, the original data is projected onto a disentangled latent space and generates new samples by combining disentangled factors from different domains. The presented experimental results showcase the contribution of the geometry-based method to cardiac image segmentation through boosting the domain generalisation, while also demonstrating the contribution of the disentangled factors mixing method to the domain adaptation. Then, I presented a novel semi-supervised meta-learning framework for semi-supervised domain generalisation. Using disentanglement the proposed approach models domain shifts, and thanks to the reconstruction approach to disentanglement, the proposed model can be trained also with unlabeled data. By applying the designed constraints (including the low-rank regularisation) to the gradient-based meta-learning approach, the model extracts robust anatomical features useful for predicting segmentation masks in a semi-supervised manner. Extensive quantitative results, especially when insufficient annotated data are available, indicate remarkable improvements compared to previous state-of-the-art approaches.

In Chapter 6, I have presented that using compositional equivariance as an inductive bias helps to learn generalisable and interpretable compositional representations. In particular, I used different learning biases in different settings to constrain the representations to be compositionally equivariant. For the unsupervised setting and weakly supervised setting, we observed that the representations achieve a certain level of compositional equivalence, which is partially interpretable. For the semi-supervised settings, I qualitatively showed that some of the representations are well interpretable when little supervision is given. Quantitatively, vMFNet and vMFPseudo, the models built based on decomposing the compositional representations with different design biases and learning biases, achieved the best generalisation performance compared to several strong baselines. Overall, as I discussed in Section 6.3 and demonstrated with the results, different learning settings and biases allow the model to learn the representations that are compositionally equivariant at different levels. I conclude that strong prior knowledge (e.g. weak supervision "*heart or not*") or some supervision significantly boosts the ability to achieve compositional equivariance. Taking advantage of the unlabeled data also plays a key role to learn compositionally equivariant representations as it implicitly constructs more groups of data that have shared factors.

The broader significance of the work is over two aspects: **a)** a new problem setting – semi-

supervised domain generalisation; **b)** a new theory for learning representations in the medical domain – compositional equivariance. As discussed in Chapter 5, I proposed a solution to the problem of domain-generalised medical image segmentation with unlabeled data. In fact, it is the first work that considers solving such a problem. Apart from the proposed methods in Chapter 6, there are already several follow-up works such as [233]. Solving the problem of semi-supervised domain generalisation for medical image segmentation is extremely challenging but also very practically applicable. I believe that future works will be inspired by our solution of combining meta-learning and disentanglement to eventually find the most reliable solution. The compositional equivariance theory opens the door to learning generalisable and interpretable representations for medical image analysis by incorporating compositionality. Considering compositional equivariance as an inductive bias, researchers can focus on devising learning and design biases with prior and expert knowledge. Achieving compositional equivariance provides more guarantee of generalisation and interpretability.

7.2 Limitations and Opportunities

In this section, I identify the limitations of the proposed methods and discuss ideas and opportunities for improvement.

In Chapter 4, I reported that a sweet spot between task performance and disentanglement needs to be achieved. However, it is still an unanswered question on how to achieve such sweet spot. It is possible to exploit the metrics to improve disentanglement itself in an iterative manner, as Esterman et al. [244] have done, while simultaneously monitoring task performance when improving disentanglement. As studied in [245], the distance correlation metric can be used as a learning objective to train the model to learn more disentangled representations, which has not been widely tested in the medical domain, remaining as an opportunity. In terms of the proposed IOB metric, incorporating it as a training objective may cause memory issues as additional networks are required and the metric performs on the whole training dataset. One may consider to exploit the batch-data based IOB objective. The moving average over multiple data batches of IOB during training can be one possible solution. On the other hand, it is unclear if the proposed metrics generalise to vector-form disentangled representations. Moreover, the proposed metrics may fail when there exist correlated factors of variation in the data. As experiments in [41] show, most existing metrics struggle when measuring the disentanglement of models trained with data that include correlated factors of variation. For

the correlated factors, one may consider a correlation calibration term that is based on the prior knowledge of the data to regularise the DC and IOB metrics. For example, for IOB, an additional correlation calibration network can be trained to un-bias the correlation.

For the Factor-based Augmentation approach presented in Chapter 5, only mixing the anatomy and modality is performed. This approach approximates the scenario that the same patients are scanned in different hospitals. It creates more combinations of anatomy and modality but does not introduce more diversity in the anatomy and modality. To create plausible anatomy, I and the co-authors investigated augmenting the anatomy by doing anatomy arithmetic in [246]. We combine disentangled anatomical factors of different input images, to create new plausible images with desired characteristics. We showed that these generated images and accompanied metadata can be used to augment existing data for improved performance. This approach improves the anatomy diversity of the augmented data. However, it is still not interpolating or extrapolating the latent space. One possible solution for interpolation is linearly combining the modality factors as the new modality factor. Then adding the adversarial training mechanism to SDNet may allow the model to generate images with a plausible modality. More advanced methods of extrapolating the anatomy and modality spaces may better address this issue. As a possible solution, integrating the pre-trained generative models (e.g. pre-trained diffusion models [247]) may help on extrapolating the latent spaces. For example, one may take the SDNet anatomy channels of different patients as the input for ControlNet [248] to generate multiple medical images that may have plausible anatomical structures to better solve the anatomy arithmetic problem. In terms of the proposed semi-supervised meta-learning approach, the segmentation module is not trained when the data are not labeled, where only the reconstruction path has gradients. This possibly has a negative effect on the segmentation module as the segmentor is not trained to process the latent space of unlabelled data. One can learn from the advances in semi-supervised image segmentation to design better learning objectives to address this issue such as the cross pseudo supervision I studied in Chapter 6.

For compositionally equivariant representation learning, I have shown improved generalisation and interpretability as evidence of better compositional equivariance. However, the direct measurement of compositional equivariance is still missing. Ideally, a metric for measuring compositional equivariance that can also be leveraged as a training objective will have a significant impact on future work. The key of compositional equivariance is that the compositional equivariant representation of a generative factor should not be affected when modifying

other representations or composing it with other representations. Consider the compositionally equivariant representations of a generative model. A possible metric can consider two aspects: **a)** when fixing one representation and modifying all other representations (i.e. composing the fixed representation with different sets of other representations), the generated images should always contain the information of the fixed representation; **b)** when modifying one representation and keep other representations fixed, the generated images should always reflect the modifications/transformations on the representation. Then, it is possible to construct other forms of weak supervision such as the presence/absence of the anatomy (e.g. left ventricle) of the heart. Also, one may use weak supervision to help with the segmentation task. Finally, I have studied the un-, weakly- and semi-supervised settings for compositionally equivariant representation learning. Self-supervised learning is a missing puzzle for compositional equivariance. One may construct more groups of data that implicitly share some generative factors under contrastive learning settings [238]. In this case, the text information from health reports of medical imaging data can be leveraged for augmenting the imaging data for self-supervised learning.

7.3 Future Directions and Open Challenges

Apart from the limitations of the proposed methods, there are several open challenges existing in representation learning for medical image analysis. In this section, I discuss these opportunities and future directions.

7.3.1 New strategies for learning disentangled representations

Learning disentangled representations requires complex architectures and objective functions. Most approaches employ several loss functions and modules and, hence multiple hyperparameters. While flexibility is desirable, tuning complex systems can be difficult and it creates a barrier for further adoption of the disentanglement paradigm by the broader research community. Methods that require less hyperparameter tuning or techniques for automating this process or less complex approaches will be welcomed. Below, I discuss two possible strategies to learn disentangled representations in a simpler fashion.

Part of the content in Section 7.3.1 and Section 7.3.4 is based on :

- **Liu, X.***, Sanchez, P.*, Thermos, S.*, O'Neil, A.Q. and Tsaftaris, S.A., 2022. Learning Disentangled Representations in the Imaging Domain. Medical Image Analysis, p.102516. *Equal contribution.

Integrating self-supervised and contrastive learning. Fundamentally speaking most disentanglement approaches we reviewed here use a reconstruction approach. This may not be necessary. Recently, contrastive learning [249, 250, 251, 252, 253] has shown impressive performance for self-supervised representation learning. In particular, patch-wise contrastive learning [254] has been successfully used as an auxiliary loss function for reinforcing disentanglement [255, 256]. Additionally, Mitrovic et al. [257] and Vonkugelgen et al. [258] developed an understanding of contrastive learning from a causal perspective and argue that it can be interpreted as CSD where the representation is focusing on learning only the content, whilst developing style invariance. Methods such as MOCO [249], SimCLR [250, 251], BYOL [252], and the Barlow Twins [253] achieve this through augmentation and regularisation. Wang et al. [238] use contrastive learning for disentangling group invariant representations. Ren et al. [259] propose to discover the disentangled representations with contrasting learning at the post-hoc stage. Zimmermann et al. [260] have taking it a step further to suggest that contrastive learning under certain assumptions can indeed invert the data generating process. While it is possible to learn representations that are robust (invariant) to specific interventions, it remains challenging to design augmentations and regularisations which are invariant to general interventions.

Intervention as a prior. Caselles et al. [65] suggest that a symmetry-based understanding of disentanglement can only be achieved upon interaction with an environment. To illustrate this point, Suter et al. [261] propose a disentanglement metric based on interventional robustness. Moreover, statistical independence between latent variables might not hold for real-life settings where the generating factors are correlated [42, 43]. With this intuition, Besserve et al. [262] provide a causal understanding of disentanglement in generative models based on interventions and counterfactuals. Leeb et al. [263] propose a strategy for probing the latent space of VAEs by applying interventions. Their method allows quantification of the consistency of the representation with a chosen prior as well as finding holes in the latent manifold. These works pave a new path for using interventions as a prior for disentangled representation learning.

7.3.2 Structured representation learning

In Chapter 3, I discussed the disentangled representations in a vector form, which can be learnt via different models e.g. VAEs and GANs. In Chapter 4 and Chapter 5, I explored and proposed the content-style disentanglement approaches. Eventually, in Chapter 6, I studied the compo-

sitional representations. Overall, all the studies do not consider any hierarchical structure of representations. In practice, there exist more complex structures in real data. For example, considering a dataset containing images of different animals, the factors "dog" and "cat" are the parent factors and the factors "colour of hairs", "tail length", etc., are the child factors, as illustrated in [264]. Prior work has primarily focused on developing hierarchical networks to learn hierarchically structured representations, either for simplified hierarchical structures [265] or on toy synthetic datasets [152]. Additionally, Li et al. [266] proposed to progressively learn the disentangled representations with VAEs from high- to low-levels of abstractions on the toy dataset. Wang et al. [264] have explored learning complete hierarchical structures of representations on large-scale, real-world datasets with extensive annotations of the structures and factors. Recently, causal representations consider the parent and child factors following a causal relationship [267], which is a specific form of structured representation. Overall, it is still an unsolved problem to learn the structured representations for real-world datasets, especially for medical data, without full supervision of the structures. In fact, medical imaging data typically accompanies extensive health reports that can be leveraged to learn hierarchical structures of representations from text information and subsequently learn the corresponding representations from imaging data, representing an area of great potential. On the other hand, the physical structure of the organs or anatomy inside organs can be considered for constructing the structure of representations. For example, the left ventricle, myocardium and right ventricle of human heart follow a nested structure. Similarly, the nested structure of brain tumors has been leveraged for better task performance as shown in [268]. For such nested structures, one may construct a hierarchical structure for the corresponding representations.

7.3.3 Interactive representation learning

I discussed in Chapter 3 that to learn the (disentangled) representations, we need to introduce different biases to make the representations identifiable. A novel form of introducing bias to the learning process is integrating humans into the loop i.e. introducing human interaction [236]. Particularly, we may favour incorporating human interactions into a model's latent representations e.g. to correct confounding behaviour [236]. This area of research is closely related to explanatory interactive learning (XIL) [269], which allows a learner to query the user or another information source interactively to obtain desired outputs for data points. In this interaction, the learner predicts a label for a data point, and the user corrects the learner's prediction as necessary, providing slightly improved feedback that is not necessarily optimal. XIL typically

incorporates the human user into the training loop by allowing for interaction via a model's explanations [270, 271]. Most XIL approaches interact via post-hoc explanations [264, 271]. To learn better representations, we instead should interact directly with the latent representations of a model, as studied in [236]. This line of research is also related to active learning [272, 273], where a learning algorithm can interactively query a user to label new data points with the desired outputs. Interactive representation learning has great potential in the medical domain, which can boost the ability to learn from fewer data and less annotation with the help of human interaction and significantly improves the explainability of deep models. To facilitate the integration of human interactions into learning good representations for medical data, the challenges mainly are constructing human understandable latent space and introducing proper structure to the representations as we discussed before.

7.3.4 Fair and disentangled representation learning

An important limitation is learning disentangled representation from correlated data [41]. As detailed in Section 3.2.4, real data is not *i.i.d.* and bias exists due to domain shifts. In these cases, it has been shown that factorization-based inductive biases as described in Section 3.3.1 are not enough to learn the true generating factors. These biases can have significant implications for fairness (biased towards sensitive attributes). Fairness is an important concept in machine learning whenever an algorithm tends to be biased towards sensitive attributes such as race or gender [274, 201]. Therefore, a fair model should be invariant to sensitive attributes. Developing fair algorithms is tightly related to domain generalisation as detailed in [275] and disentanglement provides a useful framework for dealing with this issues [276, 277, 278, 279].

References

[1] "MRI Scan." https://snc2dmri.weebly.com/components--functions.html. Accessed: 02/03/2023.

[2] W. J. Manning and D. J. Pennell, *Cardiovascular Magnetic Resonance*. Elsevier Health Sciences, 2018.

[3] "Wiggers diagram – wikipedia, the free encyclopedia." https://en.wikipedia.org/wiki/File:Wiggers_Diagram_2.svg. Accessed: 24/05/2023.

[4] V. M. Campello, T. Xia, X. Liu, P. Sanchez, C. Martín-Isla, S. E. Petersen, S. Seguí, S. Tsaftaris, and K. Lekadir, "Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks," *Frontiers in Cardiovascular Medicine*, p. 2693, 2022.

[5] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martín-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma, *et al.*, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge," *IEEE Transactions on Medical Imaging*, 2021.

[6] "Lecture 9 - spinal cord and spinal nerves." https://www.chegg.com/flashcards/lecture-9-spinal-cord-and-spinal-nerves-e3f39245-f2bb-4a2b-88fd-b4d6ee0f83b4/deck. Accessed: 08/03/2023.

[7] F. Prados, J. Ashburner, C. Blaiotta, T. Brosch, J. Carballido-Gamio, M. J. Cardoso, B. N. Conrad, E. Datta, G. Dávid, B. De Leener, *et al.*, "Spinal cord grey matter segmentation challenge," *NeuroImage*, vol. 152, pp. 312–329, 2017.

[8] O. Bernard, A. Lalande, C. Zotti, and et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[9] R. C. o. Petersen, "Alzheimer's disease neuroimaging initiative (ADNI)," *Neurology*, vol. 74, no. 3, pp. 201–209, 2010.

[10] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee, "Deep visual analogy-making," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[11] C. Burgess and H. Kim, "3D shapes dataset," *Available: https://github.com/deepmind/3d-shapes*, 2018.

[12] X. Liu, P. Sanchez, S. Thermos, A. Q. O'Neil, and S. A. Tsaftaris, "Learning disentangled representations in the imaging domain," *Medical Image Analysis*, p. 102516, 2022.

[13] 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, 2019.

[14] X. Liu, S. Thermos, G. Valvano, A. Chartsias, A. O'Neil, and S. A. Tsaftaris, "Measuring the biases and effectiveness of content-style disentanglement," in *Proc. British Machine Vision Conference (BMVC)*, 2021.

[15] X. Liu, S. Thermos, A. Chartsias, A. O'Neil, and S. A. Tsaftaris, "Disentangled representations for domain-generalized cardiac segmentation," in *Proc. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, pp. 187–195, Springer, 2020.

[16] X. Liu, S. Thermos, A. O'Neil, and S. A. Tsaftaris, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 307–317, Springer, 2021.

[17] X. Liu, S. Thermos, P. Sanchez, A. Q. O'Neil, and S. A. Tsaftaris, "vmfnet: Compositionality meets domain-generalised segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 704–714, Springer, 2022.

[18] C. Eastwood and C. K. I. Williams, "A framework for the quantitative evaluation of disentangled representations," *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3234–3243, 2016.

[20] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, 2016.

[21] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1096–1104, 2016.

[22] "Google's medical AI was super accurate in a lab. Real life was a different story." https://www.technologyreview.com/2020/04/27/1000658/google-medical-ai-accurate-lab-real-life-clinic-covid-diabetes-retina-disease/. Accessed: 09/02/2023.

[23] L. Zhang, X. Wang, D. Yang, T. Sanford, *et al.*, "Generalising deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[24] Y. Li, J. Chen, X. Xie, K. Ma, and Y. Zheng, "Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 614–623, Springer, 2020.

[25] J. Wang, S. Zhou, C. Fang, L. Wang, and J. Wang, "Meta corrupted pixels mining for medical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 335–345, Springer, 2020.

[26] E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, and A. P. King, "Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 413–423, Springer, 2021.

[27] "Cardiovascular specialist salary in United Kingdom." https://uk.indeed.com/career/cardiovascular-specialist/salaries. Accessed: 10/02/2023.

[28] 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.

[29] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, 2021.

[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proc. International Conference on Learning Representations (ICLR)*, 2014.

[31] 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 *Proc. International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019.

[32] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.

[33] 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 *Proc. European Conference on Computer Vision (ECCV)*, pp. 36–52, 2018.

[34] T. Xiao, J. Hong, and J. Ma, "ELEGANT: Exchanging latent encodings with GAN for transferring multiple face attributes," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 172–187, 2018.

[35] J. Lin, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo, "Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1254–1266, 2021.

[36] D. Lorenz, L. Bereska, T. Milbich, and B. Ommer, "Unsupervised part-based disentangling of object shape and appearance," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10964, 2019.

[37] P. Esser, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8857–8866, 2018.

[38] M. Sermesant, H. Delingette, H. Cochet, P. Jaïs, and N. Ayache, "Applications of artificial intelligence in cardiovascular imaging," *Nature Reviews Cardiology*, pp. 1–10, 2021.

[39] M. Llera Montero, C. JH Ludwig, R. Ponte Costa, G. Malhotra, and J. Bowers, "The role of disentanglement in generalisation," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[40] L. Schott, J. von Kügelgen, F. Träuble, P. Gehler, C. Russell, M. Bethge, B. Schölkopf, F. Locatello, and W. Brendel, "Visual representation learning does not generalize strongly within the same domain," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[41] F. Träuble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer, "On disentangled representations learned from correlated data," in *Proc. International Conference on Machine Learning (ICML)*, pp. 10401–10412, 2021.

[42] A. Dittadi, F. Träuble, F. Locatello, M. Wüthrich, V. Agrawal, O. Winther, S. Bauer, and B. Schölkopf, "On the transfer of disentangled representations in realistic settings," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[43] F. Träuble, E. Creager, N. Kilbertus, F. Locatello, A. Dittadi, A. Goyal, B. Schölkopf, and S. Bauer, "On disentangled representations learned from correlated data," in *Proc. International Conference on Machine Learning (ICML)*, pp. 10401–10412, PMLR, 2021.

[44] A. Meola, J. Rao, N. Chaudhary, M. Sharma, and S. D. Chang, "Gold nanoparticles for brain tumor imaging: A systematic review," *Frontiers in neurology*, vol. 9, p. 328, 2018.

[45] "Who website. Cardiovascular diseases." https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1. Accessed: 23/03/2023.

[46] J.-D. Schwalm, M. McKee, M. D. Huffman, and S. Yusuf, "Resource effective strategies to prevent and treat cardiovascular disease," *Circulation*, vol. 133, no. 8, pp. 742–755, 2016.

[47] V. Obas and R. S. Vasan, "The aging heart," *Clinical Science*, vol. 132, pp. 1367–1382, jul 2018.

[48] E. G. Lakatta and D. Levy, "Arterial and cardiac aging: Major shareholders in cardiovascular disease enterprises," *Circulation*, vol. 107, pp. 346–354, jan 2003.

[49] M. Steenman and G. Lande, "Cardiac aging and heart disease in humans," *Biophysical Reviews*, vol. 9, pp. 131–137, mar 2017.

[50] D. D. McManus, V. Xanthakis, L. M. Sullivan, J. Zachariah, J. Aragam, M. G. Larson, E. J. Benjamin, and R. S. Vasan, "Longitudinal tracking of left atrial diameter over the adult life course: Clinical correlates in the community," *Circulation*, vol. 121, pp. 667–674, feb 2010.

[51] K. M. Keller and S. E. Howlett, "Sex differences in the biology and pathology of the aging heart," *Canadian Journal of Cardiology*, vol. 32, pp. 1065–1073, sep 2016.

[52] A. G. Cota, "Spinal cord anatomy," *Deer's Treatment of Pain: An Illustrated Guide for Practitioners*, pp. 43–48, 2019.

[53] M. Larobina and L. Murino, "Medical image file formats," *Journal of digital imaging*, vol. 27, pp. 200–206, 2014.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[55] J. Fragemann, L. Ardizzone, X. Liu, S. A. Tsaftaris, J. Egger, and J. Kleesiek, "Review of disentanglement approaches for medical applications–towards solving the gordian knot of generative models in healthcare," *Under review*, 2022.

[56] A. Achille and S. Soatto, "Emergence of invariance and disentanglement in deep representations," *Journal of Machine Learning Research*, vol. 19, pp. 1–34, 2017.

[57] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, "Towards a definition of disentangled representations," *arXiv:1812.02230.*, 2018.

[58] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[59] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. International Conference on Machine Learning (ICML)*, pp. 2990–2999, 2016.

[60] T. Cohen, *Equivariant convolutional networks*. PhD thesis, University of Amsterdam, 2021.

[61] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv:2104.13478.*, 2021.

[62] K. Lenc and A. Vedaldi, "Understanding image representations by measuring their equivariance and equivalence," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 991–999, 2015.

[63] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv:2108.13624*, 2021.

[64] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, pp. 1–10, 2020.

[65] H. Caselles-Dupré, M. Garcia Ortiz, and D. Filliat, "Symmetry-based disentangled representation learning requires interaction with environments," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4606–4615, 2019.

[66] V. Thomas, J. Pondard, E. Bengio, M. Sarfati, P. Beaudoin, M.-J. Meurs, J. Pineau, D. Precup, and Y. Bengio, "Independently controllable features," *arXiv:1708.01289.*, 2017.

[67] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, p. 2180–2188, 2016.

[68] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the F-statistic loss," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, p. 185–194, 2018.

[69] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Annual Allerton Conference on Communication, Control and Computing*, 1999.

[70] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[71] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ICA: A unifying framework," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108, pp. 2207–2217, 2020.

[72] A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, vol. 12, no. 3, pp. 429–439, 1999.

[73] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[74] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, 2013.

[75] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[76] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[77] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1278–1286, 2014.

[78] B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen, "Discovering hidden factors of variation in deep networks," in *Proc. International Conference on Learning Representations Workshop (ICLRW)*, 2015.

[79] S. N, 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 *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[80] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[81] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue PCA directions (by accident)," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12406–12415, 2019.

[82] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," *arXiv:1804.03599.*, 2018.

[83] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[84] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

[85] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[86] B. Liu, Y. Zhu, Z. Fu, G. de Melo, and A. Elgammal, "OOGAN: Disentangling GAN with one-hot sampling and orthogonal regularization," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 4836—-4843, 2020.

[87] Y. Shen and B. Zhou, "Closed-form factorization of latent semantics in GANs," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1532–1540, 2021.

[88] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

[89] A. Gabbay and Y. Hoshen, "Demystifying inter-class disentanglement," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[90] D. Ruta, S. Motiian, B. Faieta, Z. Lin, H. Jin, A. Filipkowski, A. Gilbert, and J. Collomosse, "Aladin: all layer adaptive instance normalization for fine-grained style similarity," in *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 11926–11935, 2021.

[91] G. Valvano, A. Chartsias, A. Leo, and S. A. Tsaftaris, "Temporal consistency objectives regularize the learning of disentangled representations," in *Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART)*, pp. 11–19, 2019.

[92] H. Jiang, A. Chartsias, X. Zhang, G. Papanastasiou, S. Semple, M. Dweck, D. Semple, R. Dharmakumar, and S. A. Tsaftaris, "Semi-supervised pathology segmentation with disentangled representations," in *Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART)*, pp. 62–72, Springer, 2020.

[93] X. Liu, S. Thermos, A. Chartsias, A. O'Neil, and S. A. Tsaftaris, "Disentangled representations for domain-generalized cardiac segmentation," in *Proc. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, pp. 187–195, 2020.

[94] S. Y. Shin, S. Lee, and R. M. Summers, "Unsupervised domain adaptation for small bowel segmentation using disentangled representation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 282–292, Springer, 2021.

[95] J. Kalkhof, C. González, and A. Mukhopadhyay, "Disentanglement enables cross-domain hippocampus segmentation," *arXiv:2201.05650*, 2022.

[96] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[97] 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.

[98] 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.

[99] 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.

[100] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, *et al.*, "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[101] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4105–4113, 2017.

[102] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.

[103] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

[104] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[105] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2337–2346, 2019.

[106] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1180–1189, 2015.

[107] A. Gonzalez-Garcia, J. van de Weijer, and Y. Bengio, "Image-to-image translation for cross-domain disentanglement," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1287–1298, 2018.

[108] Q. Zhao, Z. Liu, E. Adeli, and K. M. Pohl, "Longitudinal self-supervised learning," *Medical Image Analysis*, vol. 71, p. 102051, 2021.

[109] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1013–1023, 2021.

[110] J. Huang, D. Guan, A. Xiao, and S. Lu, "FSDR: Frequency space domain randomization for domain generalization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6891–6902, 2021.

[111] F. Leeb, Y. Annadani, S. Bauer, and B. Schölkopf, "Structured representation learning using structural autoencoders and hybridization," *arXiv:2006.07796.*, 2020.

[112] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovski, O. Mastropietro, and A. Courville, "Adversarially learned inference," *Proc. International Conference on Learning Representations (ICLR)*, 2016.

[113] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *Proc. International Conference on Learning Representations (ICLR)*, 2016.

[114] S. Dash, V. N. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," in *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 915–924, 2022.

[115] X. Shen, T. Zhang, and K. Chen, "Bidirectional generative modeling using adversarial gradient estimation," *arXiv:2002.09161*, 2020.

[116] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang, "Disentangled generative causal representation learning," *arXiv:2010.02637*, 2020.

[117] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2223–2232, 2017.

[118] A. Almahairi, S. Rajeshwar, A. Sordoni, P. Bachman, and A. Courville, "Augmented CycleGAN: Learning many-to-many mappings from unpaired data," in *Proc. International Conference on Machine Learning (ICML)*, pp. 195–204, 2018.

[119] 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 *Proc. International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI)*, pp. 31–41, 2018.

[120] R. Zhang, T. Pfister, and J. Li, "Harmonic unpaired image-to-image translation," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[121] T. Xia, A. Chartsias, S. A. Tsaftaris, A. D. N. Initiative, *et al.*, "Consistent brain ageing synthesis," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 750–758, 2019.

[122] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Medical Image Analysis*, vol. 64, p. 101719, 2020.

[123] C. Li, H. Liu, C. Chen, Y. Pu, L. Chen, R. Henao, and L. Carin, "ALICE: Towards understanding adversarial learning for joint distribution matching," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[124] A. Jahanian, L. Chai, and P. Isola, "On the "steerability" of generative adversarial networks," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[125] A. Cherepkov, A. Voynov, and A. Babenko, "Navigating the GAN parameter space for semantic image editing," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[126] S. Duan, L. Matthey, A. Saraiva, N. Watters, C. Burgess, A. Lerchner, and I. Higgins, "Unsupervised model selection for variational disentangled representation learning," *International Conference on Learning Representations (ICLR)*, 2020.

[127] H. Kim and A. Mnih, "Disentangling by factorising," in *Proc. International Conference on Machine Learning (ICML)*, pp. 2649–2658, 2018.

[128] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in VAEs," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2615–2625, 2018.

[129] A. Kumar, P. Sattigeri, and A. Balakrishnan, "Variational inference of disentangled latent concepts from unlabeled observations," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[130] M.-A. Carbonneau, J. Zaidi, J. Boilard, and G. Gagnon, "Measuring disentanglement: A review of metrics," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[131] G. J. Székely, M. L. Rizzo, N. K. Bakirov, *et al.*, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[132] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. S. Kandola, "On kernel-target alignment," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 367–373, 2002.

[133] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with hilbert-schmidt norms," in *Proc. International conference on algorithmic learning theory (ALT)*, pp. 63–77, Springer, 2005.

[134] K. Li, L. Yu, S. Wang, and P.-A. Heng, "Unsupervised retina image synthesis via disentangled representation learning," in *Proc. International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI)*, pp. 32–41, 2019.

[135] J. Yang, X. Li, D. Pak, N. C. Dvornek, J. Chapiro, M. Lin, and J. S. Duncan, "Cross-modality segmentation by self-supervised semantic alignment in disentangled content space," in *Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART)*, pp. 52–61, 2020.

[136] C. Qin, B. Shi, R. Liao, T. Mansi, D. Rueckert, and A. Kamen, "Unsupervised deformable registration for multi-modal images via disentangled representations," in *Proc. International Conference on Information Processing in Medical Imaging (IPMI)*, pp. 249–261, 2019.

[137] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[138] S. Zhou, T. Xiao, Y. Yang, D. Feng, Q. He, and W. He, "GeneGAN: Learning object transfiguration and attribute subspace from unpaired data," in *Proc. British Machine Vision Conference (BMVC)*, 2017.

[139] S. Thermos, X. Liu, A. O'Neil, and S. A. Tsaftaris, "Controllable cardiac synthesis via disentangled anatomy arithmetic," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

[140] P. Esser, J. Haux, and B. Ommer, "Unsupervised robust disentangling of latent characteristics for image synthesis," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2699–2709, 2019.

[141] C. I. Bercea, B. Wiestler, D. Rueckert, and S. Albarqouni, "FedDis: Disentangled federated learning for unsupervised brain pathology segmentation," *arXiv:2103.03705.*, 2021.

[142] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[143] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P. Burgess, M. Bošnjak, M. Shanahan, M. Botvinick, D. Hassabis, and A. Lerchner, "Scan: Learning hierarchical compositional visual concepts," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[144] A. L. Yuille and C. Liu, "Deep nets: What have they ever done for vision?," *International Journal of Computer Vision*, vol. 129, pp. 781–802, 2021.

[145] M. Baroni, "Linguistic generalization and compositionality in modern artificial neural networks," *Philosophical Transactions of the Royal Society B*, vol. 375, no. 1791, p. 20190307, 2020.

[146] J. Andreas, "Measuring compositionality in representation learning," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[147] J. A. Fodor and E. Lepore, *The compositionality papers*. Oxford University Press, 2002.

[148] I. Mordatch and P. Abbeel, "Emergence of grounded compositional language in multi-agent populations," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[149] A. Kortylewski, J. He, Q. Liu, and A. L. Yuille, "Compositional convolutional neural networks: A deep architecture with innate robustness to partial occlusion," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8940–8949, 2020.

[150] A. Volokitin, E. Konukoglu, and L. Van Gool, "Decomposing image generation into layout prediction and conditional synthesis," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 372–373, 2020.

[151] S. Mo, M. Cho, and J. Shin, "Instagan: Instance-aware image-to-image translation," in *Proc. International Conference on Learning Representations (ICLR)*, 2018.

[152] K. K. Singh, U. Ojha, and Y. J. Lee, "Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery," in *Proc. IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 6490–6499, 2019.

[153] P. Tokmakov, Y.-X. Wang, and M. Hebert, "Learning compositional representations for few-shot recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6372–6381, 2019.

[154] T. Klinger, D. Adjodah, V. Marois, J. Joseph, M. Riemer, A. Pentland, and M. Campbell, "A study of compositional generalization in neural models," *arXiv:2006.09437*, 2020.

[155] E. Akyürek, A. F. Akyürek, and J. Andreas, "Learning to recombine and resample data for compositional generalization," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[156] A. Kortylewski, Q. Liu, H. Wang, Z. Zhang, and A. Yuille, "Combining compositional models and deep networks for robust object classification under occlusion," in *Proc. IEEE/CVF winter conference on applications of computer vision (CVPR)*, pp. 1333–1341, 2020.

[157] D. Huynh and E. Elhamifar, "Compositional zero-shot learning via fine-grained dense feature composition," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 19849–19860, 2020.

[158] N. Liu, S. Li, Y. Du, J. Tenenbaum, and A. Torralba, "Learning to compose visual relations," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021.

[159] D. Arad Hudson and L. Zitnick, "Compositional transformers for scene generation," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[160] X. Yuan, A. Kortylewski, *et al.*, "Robust instance segmentation through reasoning about multi-object occlusion," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11141–11150, 2021.

[161] Y. Zhang, A. Kortylewski, Q. Liu, *et al.*, "A light-weight interpretable compositional-network for nuclei detection and weakly-supervised segmentation," *arXiv:2110.13846*, 2021.

[162] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," in *arXiv:1210.5474*, 2012.

[163] T. S. Cohen and M. Welling, "Learning the irreducible representations of commutative lie groups," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1755–1763, 2014.

[164] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1431–1439, 2014.

[165] J. Yang, S. E. Reed, M. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3D view synthesis," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1099–1107, 2015.

[166] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, p. 2539–2547, 2015.

[167] E. Patrick, E. Sutter, and B. Ommer, "A variational U-Net for conditional appearance and shape generation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8857–8866, 2018.

[168] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "A commentary on the unsupervised learning of disentangled representations," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13681–13684, 2020.

[169] F. Träuble, E. Creager, N. Kilbertus, A. Goyal, F. Locatello, B. Schölkopf, and S. Bauer, "Is independence all you need? on the generalization of representations learned from correlated data," *arXiv:2006.07886*, 2020.

[170] Y. Xiao and W. Y. Wang, "Disentangled representation learning with Wasserstein total correlation," *arXiv:1912.12818*, 2019.

[171] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[172] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 700–708, 2017.

[173] G. Kwon and J. C. Ye, "Diagonal attention and style-based gan for content-style disentanglement in image generation and translation," in *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 13980–13989, 2021.

[174] J. Charles, T. Pfister, D. Magee, D. Hogg, and A. Zisserman., "Domain adaptation for upper body pose tracking in signed TV broadcasts," in *Proc. British Machine Vision Conference (BMVC)*, 2013.

[175] Z. Song, O. Koyejo, and J. Zhang, "Toward a controllable disentanglement network," *IEEE Transactions on Cybernetics*, 2020.

[176] X. Chang, T. Xiang, and T. M. Hospedales, "Scalable and effective deep cca via soft decorrelation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1488–1497, 2018.

[177] S. Zhou, E. Zelikman, F. Lu, A. Y. Ng, G. Carlsson, and S. Ermon, "Evaluating the disentanglement of deep generative models through manifold topology," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[178] R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole, "Weakly supervised disentanglement with guarantees," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[179] A. Papoulis and S. Unnikrishna Pillai, *Probability, random variables and stochastic processes*. McGraw-Hill: Boston, 2002.

[180] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[181] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2180–2188, 2016.

[182] X. Zhu, C. Xu, and D. Tao, "Where and what? examining interpretable disentangled representations," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[183] D. Bouchacourt, R. Tomioka, and S. Nowozin, "Multi-level variational autoencoder: Learning disentangled representations from grouped observations," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[184] T. Park, J.-Y. Zhu, O. Wang, J. Lu, E. Shechtman, A. Efros, and R. Zhang, "Swapping autoencoder for deep image manipulation," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7198–7211, 2020.

[185] E. Denton and V. Birodkar, "Unsupervised learning of disentangled representations from video," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4414–4423, 2017.

[186] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, "Decomposing motion and content for natural video sequence prediction," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[187] J.-T. Hsieh, B. Liu, D.-A. Huang, L. F. Fei-Fei, and J. C. Niebles, "Learning to decompose and disentangle representations for video prediction," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 517–526, 2018.

[188] X. Xing, T. Han, R. Gao, S.-C. Zhu, and Y. N. Wu, "Unsupervised disentangling of appearance and geometry by deformable generator network," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10354–10363, 2019.

[189] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 465–476, 2017.

[190] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv:1308.3432*, 2013.

[191] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[192] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv:1607.08022*, 2016.

[193] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.

[194] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

[195] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[196] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons," *Royal Danish Academy of Sciences and Letters*, vol. 5, no. 4, pp. 1–34, 1948.

[197] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

[198] F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," in *Proc. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, pp. 120–129, Springer, 2017.

[199] C. Chen, C. Qin, H. Qiu, *et al.*, "Deep learning for cardiac image segmentation: A review," *Frontiers in Cardiovascular Medicine*, vol. 7, no. 25, pp. 1–33, 2020.

[200] Q. Tao, W. Yan, Y. Wang, E. H. Paiman, Shamonin, *et al.*, "Deep learning–based method for fully automatic quantification of left ventricle function from cine mr images: a multivendor, multicenter study," *Radiology*, vol. 290, no. 1, pp. 81–88, 2019.

[201] E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, and A. P. King, "Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 413–423, Springer, 2021.

[202] C. Bian, C. Yuan, J. Wang, M. Li, X. Yang, S. Yu, K. Ma, J. Yuan, and Y. Zheng, "Uncertainty-aware domain alignment for anatomical structure segmentation," *Medical Image Analysis*, vol. 64, p. 101732, 2020.

[203] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, *et al.*, "Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan," *NeuroImage*, vol. 208, p. 116450, 2020.

[204] B. E. Dewey, L. Zuo, A. Carass, *et al.*, "A disentangled latent space for cross-site mri harmonization," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 720–729, Springer, 2020.

[205] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[206] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalisation via invariant feature representation," in *Proc. International Conference on Machine Learning (ICML)*, pp. 10–18, PMLR, 2013.

[207] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 624–639, 2018.

[208] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalisation by solving jigsaw puzzles," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2229–2238, 2019.

[209] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao, "Domain generalization via entropy regularization," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[210] H. Li, Y. Wang, R. Wan, S. Wang, *et al.*, "Domain generalisation for medical imaging classification with linear-dependency regularization," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[211] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalisation with adversarial feature learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5400–5409, 2018.

[212] Q. Dou, D. C. Castro, K. Kamnitsas, and B. Glocker, "Domain generalisation via model-agnostic learning of semantic features," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[213] Q. Liu, Q. Dou, and P.-A. Heng, "Shape-aware meta-learning for generalising prostate mri segmentation to unseen domains," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 475–485, Springer, 2020.

[214] X. Li, L. Yu, Y. Jin, C.-W. Fu, L. Xing, and P.-A. Heng, "Difficulty-aware meta-learning for rare disease diagnosis," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 357–366, Springer, 2020.

[215] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalisation," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1446–1455, 2019.

[216] P. Khandelwal and P. Yushkevich, "Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging," in *Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART)*, pp. 73–84, Springer, 2020.

[217] H. Sharifi-Noghabi, H. Asghari, N. Mehrasa, and M. Ester, "Domain generalisation via semi-supervised meta learning," *arXiv:2009.12658*, 2020.

[218] O. Ronneberger, P. Fischer, and T. Brox, "UNet: Convolutional networks for biomedical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, Springer, 2015.

[219] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. International Conference on Machine Learning (ICML)*, p. 807–814, 2010.

[220] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE/CVF International Conference on Computer Vision (CVPR)*, pp. 2980–2988, 2017.

[221] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 240–248, Springer, 2017.

[222] C. F. Baumgartner, L. M. Koch, M. Pollefeys, and E. Konukoglu, "An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation," in *Proc. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM)*, pp. 111–119, 2017.

[223] X. Yu, Y. Chen, T. Li, S. Liu, and G. Li, "Multi-mapping image-to-image translation via learning disentanglement," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[224] M. Ilse, J. M. Tomczak, C. Louizos, and M. Welling, "DIVA: Domain invariant variational autoencoders," in *Proc. International Conference on Medical Imaging with Deep Learning (MIDL)*, pp. 322–348, PMLR, 2020.

[225] W.-D. K. Ma, J. Lewis, and W. B. Kleijn, "The hsic bottleneck: Deep learning without back-propagation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5085–5092, 2020.

[226] A. Antoniou, H. Edwards, and A. Storkey, "How to train your MAML," in *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[227] F. Isensee, P. F. Jaeger, *et al.*, "nnUNet: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[228] M.-P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Proc. International Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 566–568, IEEE, 1994.

[229] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4080–4090, 2017.

[230] J. Tubiana and R. Monasson, "Emergence of compositional representations in restricted boltzmann machines," *Physical review letters*, vol. 118, no. 13, p. 138301, 2017.

[231] C. Chen, K. Hammernik, C. Ouyang, C. Qin, W. Bai, and D. Rueckert, "Cooperative training and latent space data augmentation for robust medical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 149–159, Springer, 2021.

[232] R. Gu, J. Zhang, R. Huang, W. Lei, G. Wang, and S. Zhang, "Domain composition and attention for unseen-domain generalizable medical image segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 241–250, Springer, 2021.

[233] H. Yao, X. Hu, and X. Li, "Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[234] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in *Proc. International Conference on Machine Learning (ICML)*, pp. 5338–5348, PMLR, 2020.

[235] A. Stone, H. Wang, M. Stark, Y. Liu, D. Scott Phoenix, and D. George, "Teaching compositionality to CNNs," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5058–5067, 2017.

[236] W. Stammer, M. Memmel, P. Schramowski, and K. Kersting, "Interactive disentanglement: Learning concepts by interacting with their prototype representations," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10317–10328, 2022.

[237] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen, "Weakly-supervised disentanglement without compromises," in *Proc. International Conference on Machine Learning (ICML)*, pp. 6348–6359, PMLR, 2020.

[238] T. Wang, Z. Yue, J. Huang, Q. Sun, and H. Zhang, "Self-supervised learning disentangled group representation as feature," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[239] F. Milletari, N. Navab, and S.-A. Ahmadi, "VNet: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[240] X. Chen, Y. Yuan, G. Zeng, and J. Wang, "Semi-supervised semantic segmentation with cross pseudo supervision," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622, 2021.

[241] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[242] Y. He, A. Carass, L. Zuo, *et al.*, "Autoencoder based self-supervised test-time adaptation for medical image analysis," *Medical Image Analysis*, vol. 72, p. 102136, 2021.

[243] G. Valvano, A. Leo, S. A. Tsaftaris, *et al.*, "Re-using adversarial mask discriminators for test-time training under distribution shifts," *Machine Learning for Biomedical Imaging*, vol. 1, pp. 1–10, 2022.

[244] B. Estermann, M. Marks, and M. F. Yanik, "Robust disentanglement of a few factors at a time," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13387–13398, 2020.

[245] X. Zhen, Z. Meng, R. Chakraborty, and V. Singh, "On the versatile uses of partial distance correlation in deep learning," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 327–346, Springer, 2022.

[246] S. Thermos, X. Liu, A. O'Neil, and S. A. Tsaftaris, "Controllable cardiac synthesis via disentangled anatomy arithmetic," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 160–170, Springer, 2021.

[247] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

[248] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv:2302.05543*, 2023.

[249] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2020.

[250] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

[251] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 22243–22255, 2020.

[252] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Advances in neural information processing systems (NeurIPS)*, vol. 33, pp. 21271–21284, 2020.

[253] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. International Conference on Machine Learning (ICML)*, pp. 12310–12320, PMLR, 2021.

[254] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 319–345, 2020.

[255] L. Zhou, J. Bae, H. Liu, G. Singh, J. Green, D. Samaras, and P. Prasanna, "Chest radiograph disentanglement for COVID-19 outcome prediction," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 345–355, Springer, 2021.

[256] D. Tomar, L. Zhang, T. Portenier, and O. Goksel, "Content-preserving unpaired translation from simulated to realistic ultrasound images," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 659–669, 2021.

[257] J. Mitrovic, B. McWilliams, J. C. Walker, L. H. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[258] J. V. Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello, "Self-supervised learning with data augmentations provably isolates content from style," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[259] X. Ren, T. Yang, Y. Wang, and W. Zeng, "Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[260] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel, "Contrastive learning inverts the data generating process," in *Proc. International Conference on Machine Learning (ICML)*, pp. 12979–12990, PMLR, 2021.

[261] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *Proc. International Conference on Machine Learning (ICML)*, pp. 6056–6065, 2019.

[262] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, "Counterfactuals uncover the modular structure of deep generative models," *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[263] F. Leeb, S. Bauer, and B. Schölkopf, "Interventional assays for the latent space of autoencoders," *arXiv:2106.16091*, 2021.

[264] A. Wang, W.-N. Lee, and X. Qi, "Hint: Hierarchical neuron concept explainer," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10254–10264, 2022.

[265] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. Burgess, M. Bošnjak, M. Shanahan, M. Botvinick, D. Hassabis, and A. Lerchner, "Scan: Learning hierarchical compositional visual concepts," in *Proc. International Conference on Learning Representations (ICLR)*, vol. 6, 2018.

[266] Z. Li, J. V. Murkute, P. K. Gyawali, and L. Wang, "Progressive learning and disentanglement of hierarchical representations," in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[267] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris, "Causal machine learning for healthcare and precision medicine," *Royal Society Open Science*, vol. 9, no. 8, p. 220638, 2022.

[268] C. Reddy, K. Gopinath, and H. Lombaert, "Brain tumor segmentation using topological loss in convolutional networks," in *Proc. International Conference on Medical Imaging with Deep Learning (MIDL)*, 2019.

[269] S. Teso and K. Kersting, "Explanatory interactive machine learning," in *Proc. AAAI/ACM Conference on AI, Ethics, and Society*, pp. 239–245, 2019.

[270] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting, "Making deep neural networks right for the right scientific reasons by interacting with their explanations," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 476–486, 2020.

[271] W. Stammer, P. Schramowski, and K. Kersting, "Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations," in *Proc. IEEE/CVF conference on computer vision and pattern recognition*, pp. 3619–3629, 2021.

[272] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *Proc. International conference on machine learning (ICML)*, pp. 1183–1192, PMLR, 2017.

[273] S. Hanneke *et al.*, "Theory of disagreement-based active learning," *Foundations and Trends® in Machine Learning*, vol. 7, no. 2-3, pp. 131–309, 2014.

[274] E. Puyol-Antón, B. Ruijsink, J. Mariscal Harana, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi, P. Chowienczyk, and A. P. King, "Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation," *Frontiers in cardiovascular medicine*, p. 664, 2022.

[275] E. Creager, J.-H. Jacobsen, and R. Zemel, "Environment inference for invariant learning," in *Proc. International Conference on Machine Learning (ICML)*, pp. 2189–2200, PMLR, 2021.

[276] F. Locatello, G. Abbati, T. Rainforth, S. Bauer, B. Schölkopf, and O. Bachem, "On the fairness of disentangled representations," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[277] E. Creager, D. Madras, J.-H. Jacobsen, M. Weis, K. Swersky, T. Pitassi, and R. Zemel, "Flexibly fair representation learning by disentanglement," in *Proc. International Conference on Machine Learning (ICML)*, pp. 1436–1445, PMLR, 2019.

[278] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 746–761, Springer, 2020.

[279] L. Xianjing, B. E. E., N. W. J., W. E. B., R. G. V, and Li, "Projection-wise disentangling for fair and interpretable representation learning: Application to 3d facial shape analysis," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 814–823, 2021.