University of London
Imperial College of Science, Technology and Medicine
Department of Computing

Towards Autonomous Diagnostic Systems with Medical Imaging

Athanasios Vlontzos

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of the University of London and
the Diploma of Imperial College, June 2022

Declaration

I hereby declare that the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgments.

Athanasios Vlontzos

June 2022

Abstract

Democratizing access to high quality healthcare has highlighted the need for autonomous diagnostic systems that a non-expert can use. Remote communities, first responders and even deep space explorers will come to rely on medical imaging systems that will provide them with Point of Care diagnostic capabilities.

This thesis introduces the building blocks that would enable the creation of such a system. Firstly, we present a case study in order to further motivate the need and requirements of autonomous diagnostic systems. This case study primarily concerns deep space exploration where astronauts cannot rely on communication with earth-bound doctors to help them through diagnosis, nor can they make the trip back to earth for treatment. Requirements and possible solutions about the major challenges faced with such an application are discussed.

Moreover, this work describes how a system can explore its perceived environment by developing a Multi Agent Reinforcement Learning method that allows for implicit communication between the agents. Under this regime agents can share the knowledge that benefits them all in achieving their individual tasks. Furthermore, we explore how systems can understand the 3D properties of 2D depicted objects in a probabilistic way.

In Part II, this work explores how to reason about the extracted information in a causally enabled manner. A critical view on the applications of causality in medical imaging, and its potential uses is provided. It is then narrowed down to estimating possible future outcomes and reasoning about counterfactual outcomes by embedding data on a pseudo-Riemannian manifold and constraining the latent space by using the relativistic concept of light cones.

By formalizing an approach to estimating counterfactuals, a computationally lighter alternative to the *abduction-action-prediction* paradigm is presented through the introduction of Deep Twin Networks. Appropriate partial identifiability constraints for categorical variables are derived and the method is applied in a series of medical tasks involving structured data, images and videos.

All methods are evaluated in a wide array of synthetic and real life tasks that showcase their abilities, often achieving state-of-the-art performance or matching the existing best performance while requiring a fraction of the computational cost.

Acknowledgements

This thesis would not have been possible without the inputs of a wide array of people. I would like to thank first and foremost my parents for their continual support and advice, without them nothing like this would be possible. Moreover, I would like to thank my advisors Bernhard Kainz and Daniel Rueckert for their guidance and support and for allowing me to investigate some admittedly fringe topics in machine learning.

It would be impossible to forget the contributions of my collaborators and friends including but not limited to Henrique Bergallo Rocha, Haydrien Reynaud, Luca Schmidtke, Sam Budd, Amir Alansary, Benjamin Hou, Ciarán Gilligan-Lee, Asti Bhat and many many more. But a PhD journey is not only work - it also includes discussions about politics, economics, gastronomy, movies, music and Friday 5pm pints, for that I need to give special mention to Miguel Monteiro, Margherita Rosnati, Nick Pawlowski, Charlie Jones, Gabe Sutherland, Theo Franquet, Youssef Rizk, Ilia Kokkali, Ioannis Giannakoulis, Giannis Nikiteas, Thomas Rarris, Ezer Moysis, John Tsipouras, Vasiliki Kalogianni and Angelos Filos.

A toast to future collaborations and happy times.

Dedication

To my parents.

J.& L.

"Today's Utopia is Tomorrow's Reality"

A. V.

"I have come to believe that the whole world is an enigma, a harmless enigma that is made terrible by our own mad attempt to interpret it as though it had an underlying truth."

"Then why do you want to know?"

"Because learning does not consist only of knowing what we must or we can do, but also of knowing what we could do and perhaps should not do."

Umberto Eco - Foucault's Pendulum & The Name Of The Rose

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Objectives

Throughout human history, people have gazed up to the stars above and imagined, how they would look like, how would they travel between them exploring new worlds and coming into contact with new civilizations. We currently live at the start of the space age of humanity. Approximately 50 years ago the first man set foot on the moon while currently there are discussions and proposed timelines for manned expeditions to Mars. As the human species travels further away from the safety of our home planet new challenges have to be overcome, these are not only concerning spaceships and propulsion drives but most importantly the continual well-being of humans.

We are fragile beings. We cannot withstand major fluctuations in radiation, temperature, pressure or consistency of the atmosphere. We cannot survive without food and water. Our evolution and development has been dictated by the constant presence of 1g gravity. As we travel away from the protective enclave of our atmosphere we will be faced with changes in our anatomy and the development of new pathologies that we cannot observe on Earth.

Greater distances also mean greater solitude. Limited by fundamental laws of physics we are unable to have instantaneous travel and communications. Tele-medicine and evacuation to the

safety of Earth become infeasible as the distances grow. As such the first explorers of the outer solar system (beyond the asteroid belt, between Mars and Jupiter) will have to be self-reliant for most if not all their well-being needs.

In this Thesis we will be looking into the basic building blocks required to build such advanced systems that are able to take care of humans in their explorations of the Cosmos.

Space exploration is a principle example of "Big Science". A large scale project that strives for an extremely ambitious but crucial target outcome. These projects usually funded by military and governmental budgets have a major, sometimes overlooked, advantage. They tend to have significant scientific discoveries and technologies as by-products of their main development path. In the case above, where humans try to develop an autonomous diagnostic system able to perform its duties in the vastness of interplanetary space, we stand to gain great technological advances for the people that will remain on our planet.

Even before achieving the status of "space capable" an autonomous diagnostic system can aid remote communities by performing triage and initial diagnosis, saving crucial time in the treatment of the residents. It can assist first responders in the scene of a natural disaster by giving them the tools to assist more people in a shorter period of time. Finally it can even help medical personnel in large hospitals by automating monitoring and diagnostic procedures so doctors and nurses can focus on the time-sensitive cases.

We are interested in the building blocks that would enable the creation and deployment of such an autonomous intelligent system. We recognize the complexity and difficulty of the overall task hence we will focus on key points that it is the belief of the author that they are crucial for any realization of an ADS.

In Figure 1.1 we identify some of the concepts and building blocks required in the efforts to build an ADS. We will be referring back to this graphic in order to help the reader relate the concepts introduced throughout the thesis.

Straight away we can identify three main input types we could be using in an ADS. *Vision* refers to any form of visual input, images or videos that can serve as an information source

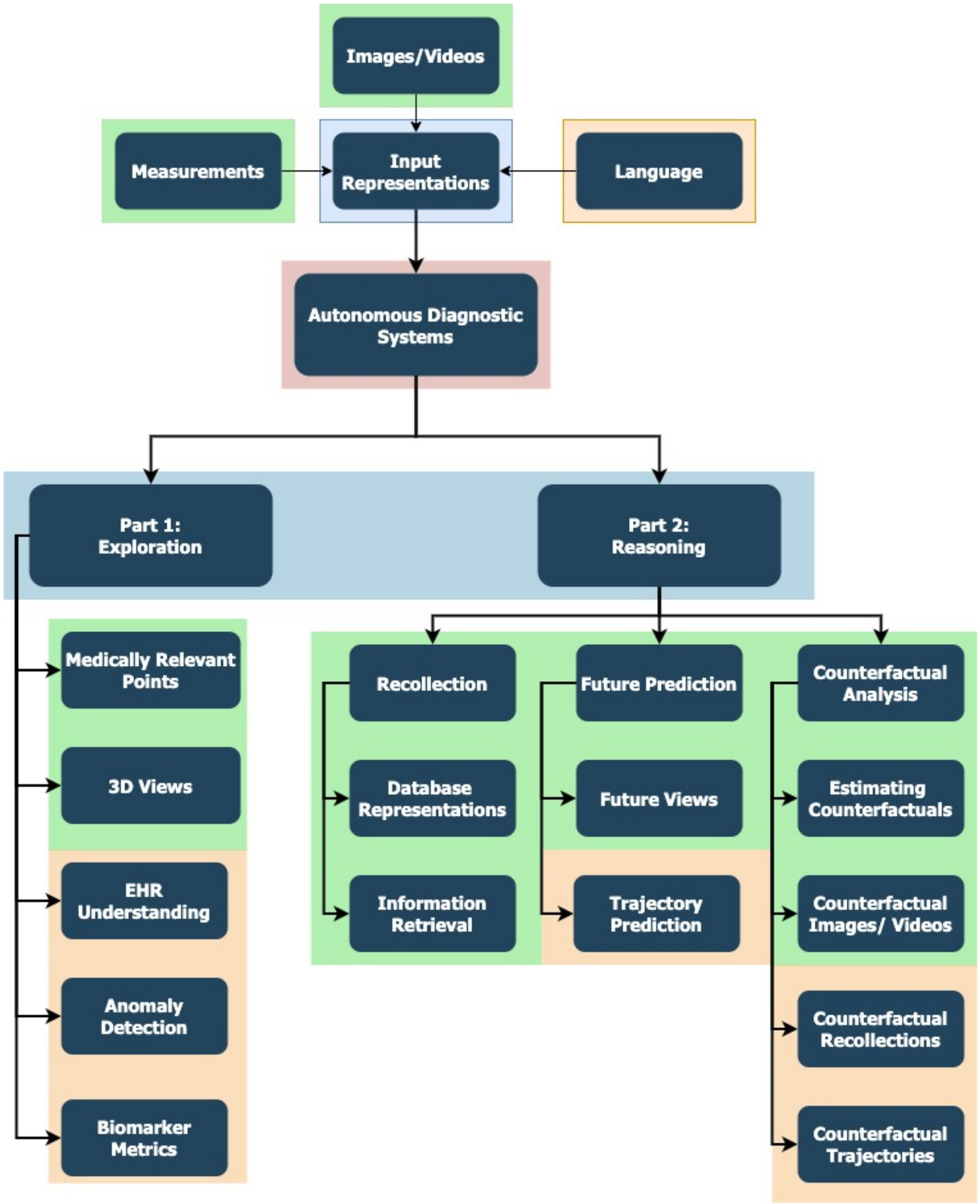


Figure 1.1: Block diagram of key components of an Autonomous Diagnostic System. This figure serves as guidance for the elements that make up this thesis. Green background blocks have been addressed and exist in this thesis; Orange background blocks are considered future work. The left hand side – *Exploration* – serves as an information extractor while the right hand side – *Reasoning* – serves as the reasoning required for any diagnosis.

for our diagnostic system. Examples that have been used in the subsequent chapters include Ultrasound (US), Magnetic Resonance Imaging (MRI) and Computed Tomography (CT). Visual inputs can aid our system with information about real time conditions and state of the patient's anatomy. However these can be augmented with a variety of measured biosignals like Electroencephalographs (EEG), Electrocardiograph (ECG), Oxymeters, blood tests and many others. These inputs allow us to gain a deeper understanding of the real time and past states of the patient and can inform greatly both diagnosis and treatment planning. Finally linguistic inputs like Electronic Health Records (EHR) or self reporting of symptoms can guide diagnostic systems to prune down possible diagnosis and arrive, hence, in a more accurate estimation of the pathology.

In Figure 1.1 there is a distinction between two main concepts - *Exploration* and *Reasoning*. In the *Exploration* side we group all methods and components that are primarily concerned with information extraction from the input sources. That can include identifying and localizing medically relevant points like anatomical landmarks - detailed in Chapter 4. It also includes creating accurate models of the scenes the system observes, imagining, hence, how the object in its field of vision appears in the real world - detailed in Chapter 5. Moreover it is believed that crucial parts of an ADS are the understanding of Electronic Health Records (EHR); detecting anomalies from all the given inputs and performing extra measurements of biomarkers, commonly used in medical practice. The latter three components are not the concern of this thesis and are left as some very interesting and challenging future work.

Moreover, in an effort to increase our diagnostic capabilities we need to think of the potential outcomes of our actions, treatments and disease progression. In Chapter 6 and Chapter 8 we are investigating the production of future views of our instruments from a causal perspective. Establishing causal limitations allows us to have faith in our system that the output will be consistent with the laws of medicine as we understand them and that it will not generate nonsensical images and videos.

Finally any diagnostic system must be able to reason about potentially different outcomes in terms of cause and effect in order to compare them with the observed measurements and draw

accurate conclusions. For example, a doctor might encounter the following situation: given the patient's blood tests if the patient had leukemia the blood tests should be showing a increase of white blood cells, but this is not observed in reality, hence the symptoms experienced by the patient cannot be attributed to the diagnosis of leukemia. Reasoning in a fashion that includes cause and effect allows a system to transcend from the probabilistic guessing paradigm and enter the formal causality one, thus increasing the trust we can have in autonomous diagnostic systems.

In Chapter 7 we develop novel methods for partially identifiable computationally lightweight counterfactual inference that admit questions like "Given that we gave the patient drug A and the symptoms did not get better, had we given them drug B would they have had a better outcome?". We focus on some of the most crucial questions in diagnosis: the probabilities of Necessity, Sufficiency and Necessity and Sufficiency for an outcome given a treatment. These are the most important questions a doctor asks implicitly or explicitly in the course of a diagnosis. Counterfactual trajectories of patients and recollection of data are left for future work.

1.2 Contributions

The main contributions of the thesis are in greater detail:

- Chapter 2 establishes the required theoretical background
- Chapter 3 presents a case study. In a deep space exploration setting we argue about the necessity of autonomous point of care diagnostic systems and highlight the desiderata of such a system both in terms of hardware and software. We furthermore, look into possible ways of tackling the main challenges present in this use case.
- Chapter 4 introduces the concepts of multiple agents exploring an environment as defined by a medical volume. In this section we develop a novel method of implicit communications among the agents that attempt to locate medical landmarks like the Anterior Commissure (AC) and the Posterior Commissure (PC). We evaluate our method in MRI

and US volumes of the brain and the heart respectively achieving state-of-the-art performance. We hence establish a method of exploration for the agent in order to identify medically significant points in the scanned volumes.

- Chapter 5 provides a novel Deep Learning (DL) methodology that tackles the ill-defined problem of 2D to 3D reconstruction and segmentation. Given a 2D scan we build a neural network capable of reconstructing possible 3D volumes and segments to extract characteristics of the anatomy. With this contribution we seek to establish a method to imagine how objects look like in the 3D world by only observing 2D projections. We build a fully probabilistic approach to overcome the loss of information that characterized the original creation of the 2D projection of the object from 3D.

- Part II represents the bulk of the work done towards intelligent autonomous diagnostic agents. In Chapters 6 and 7 we introduce two novel DL methods. First we present a novel approach that combines special relativity [Ein15] with hyperbolic geometry Machine Learning (ML) in order to create an algorithm able to generate plausible causally related future outcomes from videos. Secondly, we create the DL equivalent of Balke and Pearl's [BP94] Twin Network by proposing the Deep Twin Networks, able to perform accurate and computationally lighter counterfactual inference. We then showcase the abilities of our Deep Twin Networks by generating realistic counterfactual US videos in Chapter 8.

- The penultimate topic in Chapter 9 explores an auxiliary concept of how we can leverage causality to efficiently train a Machine Learning (ML) method under significant data constraints.

1.3 Publications

The chapters of this thesis follow a series of nine published papers and papers under review (as of the time of writing). All are first authored papers with the exception of one which is co-first authored. This is followed by a list of further first, co-first and co authored papers that are auxiliary to the concepts mentioned in this thesis.

Published:

- **Vlontzos A**, Alansary A, Kamnitsas K, Rueckert D, Kainz B. Multiple landmark detection using multi-agent reinforcement learning. In International conference on medical image computing and computer-assisted intervention (MICCAI) 2019 Oct 13 (pp. 262-270). Springer, Cham. **Contributions:** A.V. conceived of the study, developed the appropriate tools and methods, ran the experiments and led the writing of the paper. A.A. supported the development of methods and tools and contributed to the manuscript. K.K. conceived of part of the ablation study. D.R. and B.K. supervised the findings of this work and contributed to the final manuscript.

- **Vlontzos A**, Budd S, Hou B, Rueckert D, Kainz B. 3D probabilistic segmentation and volumetry from 2D projection images. In International Workshop on Thoracic Image Analysis 2020 Oct 8 (pp. 48-57). Springer, Cham. **Contributions:** A.V. and B.H conceived and planned the experiments. A.V. and S.B. ran the experiments. D.R. and B.K. supervised the project. All authors contributed to finalizing the manuscript.

- **Vlontzos A**, Sutherland G, Ganju S, Soboczenski F. Next-Gen Machine Learning Supported Diagnostic Systems for Spacecraft. AI for Spacecraft Longevity, IJCAI workshop. 2021 Jun 10. **Contributions:** A.V. G. Suth. and F.S. conceived the case study and set out the key elements. All authors contributed to finalizing the manuscript.

- H. Reynaud, **A. Vlontzos**, M. Dombrowski, C. M. Gilligan-Lee, A. Beqiri, P. Leeson, and B. Kainz, "D'artagnan: Counterfactual image and video generation", In International conference on medical image computing and computer-assisted intervention (MICCAI) , 2022. **Contributions:** A.V. designed and directed the project; H.R., A.V., M.D. performed the experiments; A.V., C.M.G.L. developed the theoretical framework, A.B, P.L and B.K supervised the project; A.V., H.R., M.D, C.M.G.L. and B.K contributed to the manuscript

- **Vlontzos A**, Kainz B, Gilligan-Lee CM. Estimating categorical counterfactuals via deep twin networks. arXiv preprint arXiv:2109.01904. 2021 Sep 4. In International Confer-

ence in Machine Learning (ICML) Workshop on Beyond Bayes: Paths Towards Universal Reasoning Systems; In Uncertainty in Artificial Intellifence (UAI) Causal Inference Workshop; Under review in Nature Machine Intelligence. **Contributions:** A.V and C.M.G.L. conceived the methods and planned the experiments. A.V. developed the tools and experiments. B.K. and C.M.G.L. supervised the project. All authors contributed to finalizing the manuscript.

- **A. Vlontzos**, D. Rueckert, and B. Kainz, "A Review of Causality for Learning Algorithms in Medical Image Analysis", Journal of Machine Learning for Biomedical Imaging (MELBA) , 2022, 2022:028. **Contributions:** A.V. led the literature review and categorization of works; D.R. and B.K. supervised the project. All authors contributed to finalizing the manuscript.

Under Review & Preprints:

- **Vlontzos A**, Rocha HB, Rueckert D, Kainz B. Causal future prediction in a Minkowski space-time. arXiv preprint arXiv:2008.09154. 2020 Aug 20. Under review in Journal of Machine Learning Research (JMLR) **Contributions:** A.V. conceived, developed and run the experiments for the proposed methods and tools. H.B.R. supported the theoretical formulation; D.R. and B.K. supervised the project. All authors contributed to finalizing the manuscript.

- **Vlontzos A**, Cao Y, Schmidtke L, Kainz B, Monod A. Topological Information Retrieval with Dilation-Invariant Bottleneck Comparative Measures. arXiv preprint arXiv:2104.01672. 2021 Apr 4.; **Contributions:** A.V. devised the project, the main conceptual ideas and proof outline. A.V., Y.C and A.M developed the mathematical tools and proofs, A.V and Y.C. planned and ran the experiments. B.K. and A.M. supervised the project. All authors contributed to finalizing the manuscript.

- **A. Vlontzos**, H. Reynaud, D. Rueckert, and B. Kainz, "Is more data all you need ? a causal exploration", Under Review, 2022. **Contributions:** A.V. led the design

and implementation of the research; A.V. and H.R. ran the experiments; D.R. and B.K. supervised the project. All authors contributed to finalizing the manuscript.

Co-authored Papers; miscellaneous and relevant to the topics of this thesis papers:

- Schmidtke L, **Vlontzos A**, Ellershaw S, Lukens A, Arichi T, Kainz B. Unsupervised human pose estimation through transforming shape templates. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021 (pp. 2484-2494).

- Lamb K, Malhotra G,**Vlontzos A**, Wagstaff E, Baydin AG, Bhiwandiwalla A, Gal Y, Kalaitzis A, Reina A, Bhatt A. Prediction of GNSS phase scintillations: A machine learning approach. Machine Learning for the Physical Sciences, NeurIPS Workshop. 2019 Oct 3.

- Lamb K, Malhotra G, **Vlontzos A**, Wagstaff E, Baydin AG, Bhiwandiwalla A, Gal Y, Kalaitzis A, Reina A, Bhatt A. Correlation of auroral dynamics and GNSS scintillation with an autoencoder. Machine Learning for the Physical Sciences, NeurIPS Workshop. 2019 Oct 4.

- Reynaud H, **Vlontzos A**, Hou B, Beqiri A, Leeson P, Kainz B. Ultrasound Video Transformers for Cardiac Ejection Fraction Estimation. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2021 Sep 27 (pp. 495-505). Springer, Cham.

- G Sutherland, F Soboczenski, **A Vlontzos**, A Deep Reinforcement Learning Approach to Train Autonomous Space Debris Remediation Spacecraft, 43rd COSPAR Scientific Assembly. Held 28 January-4 February 2021

- Hou B, Vlontzos A, Alansary A, Rueckert D, Kainz B. Flexible Conditional Image Generation of Missing Data with Learned Mental Maps. InInternational Workshop on Machine Learning for Medical Image Reconstruction 2019 Oct 17 (pp. 139-150). Springer, Cham.

- Budd S, Sinclair M, Day T, **Vlontzos A**, Tan J, Liu T, Matthew J, Skelton E, Simpson J, Razavi R, Glocker B. Detecting Hypo-plastic Left Heart Syndrome in Fetal Ultrasound via Disease-Specific Atlas Maps. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2021 Sep 27 (pp. 207-217). Springer, Cham.

- Liu T, Meng Q, **Vlontzos A**, Tan J, Rueckert D, Kainz B. Ultrasound video summarization using deep reinforcement learning. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2020 Oct 4 (pp. 483-492). Springer, Cham.

- Alansary A, Oktay O, Li Y, Le Folgoc L, Hou B, Vaillant G, Kamnitsas K, **Vlontzos A**, Glocker B, Kainz B, Rueckert D. Evaluating reinforcement learning agents for anatomical landmark detection. Medical image analysis. 2019 Apr 1;53:156-64.

Chapter 2

Background

2.1 Introduction

We start off by laying out the basic concepts required to understand the subsequent chapters. We will offer a brief mention on the concepts of medical image analysis, machine learning and then we will dive deeper into the topics of relativistic physics and causality. Each section will be self sufficient such the reader can focus on themes they are unfamiliar with.

2.2 Medical Image Analysis

As humanity's interest in anatomy peaked in the Victorian Era with physicians systematizing the categorizations and descriptions of anatomical parts, we gained an immense volume of knowledge about internal medicine and diseases. The chance discovery of X-Rays by Wilhem Röntgen, Figure 2.1b – for which he received the first Nobel Prize in Physics – created a new field of medicine that allowed doctors to leverage the aforementioned knowledge gained by anatomical dissections and non invasive observations the human body to diagnose and treat patients. Since then the field has erupted to include multiple different modalities of imaging like Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Ultrasound

(US), Computed Tomography (CT), X-Rays and others. The main goals of medical imaging can thus be summarized to the non invasive observation and depiction of the human body in order to diagnose and treat the patients.

(a) Modern X-Ray by Ptrump16 - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=115012214

(b) By Wilhelm Röntgen, Public Domain, https://commons.wikimedia.org/w/index.php?curid=5059748

Figure 2.1: X-Rays (a)Modern X-Ray of a hand produced by Ptrump16 - a wikipedia editor (b). First X-Ray by Wilhelm Röntgen of his wife's hand.

Medical imaging however does have its limitations. Modalities like CT, X-Rays and Positron Emission Tomography (PET) require the patient to be exposed to ionizing radiation or even dosed with radioactive substances in order to acquire the image. Even though the doses for a typical exam are small enough not to cause any harm to the patient there are limitations on who and when can be examined under these circumstance, with pregnant women for example told to avoid PET scans. Moreover, medical personnel present in these examinations are also exposed to ionizing radiation that has cumulative effect on their well-being. As such, personnel are often being told to reduce the hours that they operate in such environments to reduce the risk and harm to themselves. On the other hand, modalities like US and MRI do not use any

Modality	Common Use	Pros	Cons
X-Rays	Bones and Some soft tissues	Inexpensive Abundant everywhere Fast	Not all tissues can be depicted Radiation exposure
CT Scans	Bones and Some soft tissues	Fast Better visual accuracy	Radiation exposure May require contrast agents Expensive
MRI	All soft tissues	No Radiation Doesn't require contrast agents Very good depiction of structures	Hard to depict moving patients or structures Expensive
US	All soft tissues	Real Time Inexpensive Portable No Radiation	Operator-dependent Image quality usually suffers

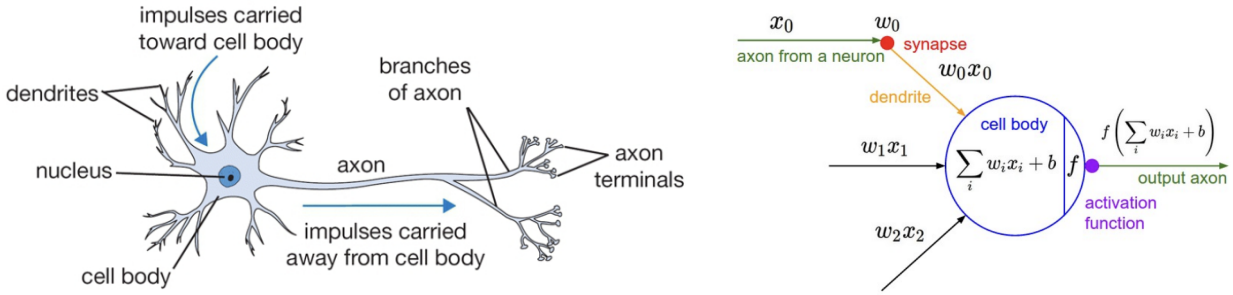
Table 2.1: Table of Advantages and Disadvantages of some of the most popular medical imaging modalities

form of ionizing radiation making them safe to use for both patients and doctors alike.

From an image quality perspective, once again there is no clear preferable modality with different methods presenting strengths in depicting different structures. For example for a quick, inexpensive view into bones one might be referred for an X-Ray scan; while for real time examinations of the fetus US is the most popular technology. Moreover, modalities like US suffer from high noise components, while X-Rays essentially depict the shadow of the body making detailed soft tissue imaging hard. MRIs tend to have the best image quality being able to depict all possible structures, however acquiring a high fidelity MRI scan is both expensive and time consuming. As such we observe that there is no single best modality, in Table 2.1, below, we have codified some of the major advantages and disadvantages of the most popular imaging techniques.

2.3 Machine Learning & Deep Learning

Artificial intelligence has fascinated humans for centuries, however, only during the past century were we able to make actual progress by developing techniques to process large quantities of data and teach algorithms to generalize. Inspired by the way neurons function in the human brain the term Artificial Neural Network (NN) was coined to describe the novel machine learning



(a) Illustration of biological neuron

(b) Artificial neuron, used in artificial neural networks

Figure 2.2: Illustrations of biological and artificial neurons. We note the similarities between the two. Image source[Kar16]

methods.

2.3.1 Artificial Neural Networks

Introduced by F. Rossenblatt in 1958 [Ros58] the Perceptron directly mimicked the method a human neuron "accumulates" and "fires" electrical discharges. In Figure 2.2 we show an illustration that showcases the parallels between a biological neuron and a "neuron" that serves as the building block of a Perceptron. Despite the promise of the Perceptron it was quickly discovered that it was limited in terms of expressivity and set of problems in could solve, as such a few modifications where made to the original model in order to increase its abilities. First a non linearity was introduced to enable the network to model complex non linear relations. Secondly a cascade of multiple, varied sized, layers of neurons were introduced in order to model hierarchical relations. Formally the output of the Perceptron can be described as

$$y = \phi(\sum_{i=1}^{m}(x_i w_i + b)) \tag{2.1}$$

where x is the input to the neuron – be that the input to the network or the output of a previous neuron – w, b is the set of weights and biases we aim to learn through our learning procedure; m signifies the number of inputs to the neuron and ϕ is the non-linearity. Common non-linearities are the sigmoid function, the hyperbolic tangent (tanh) and the rectified linear

unit (ReLu) family of functions. Non-linearites have to be differentiable and depending on their characteristic responses we aim to extract different advantages.

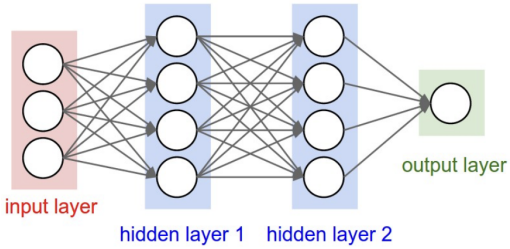


Figure 2.3: A 3-layer multi-layered Perceptron with 2 hidden layers. Image source[Kar16]

2.3.2 Convolutional Neural Networks

A Convolutional neural network (ConvNet) presents an alternative to the fully structured MLP discussed in the section above. In an effort to decrease the computational burden of relating each part of the input to all other inputs, create sparse models and share parameters between inputs, convolutions replace the linear functions of Equation (2.1). Formally defined as in Equation (2.2) convolutions allow the models to reuse the same kernels g across different part of the inputs – usually images – and extract the appropriate information. An array of 2D or sometimes 3D convolutional layers make up the backbone of ConvNets. The aforementioned kernels can then extract useful features across multiple levels and form cascades of information allowing them to build intricate and complex representations of the inputs to the model. Other common layers used in ConvNets include pooling where the input to the layer is pooled to a lower dimensionality commonly either by choosing the max or the average of the window; normalization layers which normalize the input based on its extracted statics. ConvNets also exhibit useful properties like translation equivariance *i.e.* the response of the kernel is the same regardless if the translation operation is performed before or after the convolution. ConvNets are the *de facto* standard method of extracting information from images and make up the baseline for many successful ML methods.

$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t-\tau)d\tau \tag{2.2}$$

2.3.3 Loss Functions & Optimization

In order to construct a machine learning method a training signal has to be defined such that the network "learns" from its mistakes. There is virtually infinite pool of possible loss functions one might construct or use to train a machine learning model. We characteristically mention two elementary ones that allow classification and regression respectively.

Cross-Entropy When attempting to perform a classification task a common loss function is the cross-entropy loss - binary or categorical. Measuring the difference between two probability distributions the cross entropy loss pushes the predicted from the model to match the probability distribution of the true class. Inspired by concepts of information theory and information entropy - not to be confused with thermodynamical entropy - Cross-Entropy is defined in Equation (2.3) where y_o is the true class of sample o, and $p_{i,o}$ is the predicted probability that sample o belongs to class i.

$$CE = -\sum_{i}^{C} y_o \log(p_{i,o}) \tag{2.3}$$

To aid our classification in multi-class scenarios, ground truth classes are often encoded in one-hot vectors and the output logits of the network are passed through a softmax function that normalizes them to be in the range $[0, 1]$ and add up to 1 casting them to a pseudo-probability. In the binary classification case we defined the Binary cross entropy as

$$BCE = -[y \log(p) + (1 - y) \log(1 - p)] \tag{2.4}$$

where p is once again the output of the model passed through a sigmoid function that casts the logit to $\{0, 1\}$.

Mean Squared Error The most abundant and simple loss function for regression is the mean squared error. Under this loss function we seek to minimize the euclidean distance between the point we predict and its ground truth label.

$$MSE = \frac{1}{N} \sum_{i}^{N} (y_i - \hat{y}_i)^2 \tag{2.5}$$

where $\hat{y}$ is the prediction for sample i. Despite the popularity and useful properties of the MSE loss we see in works like [LMV+19] that it is not always the appropriate loss function as it fails to capture rare events and over-smooths the output. As such other regression losses like L1, Hinge have been used depending on the task at hand.

At each chapter in this thesis we will be defining the loss function we are using. Each loss function is subsequently minimized to reduce the error rate. We normally use euclidean optimization techniques such like Stochastic Gradient Descent (SGD) seen in Equation (2.6) where ∇J is the gradient of the loss function, θ is the set of parameter of our model at time t and η the learning rate.

$$\theta_{t+1} = \theta_t - \eta \nabla J(\theta_t) \tag{2.6}$$

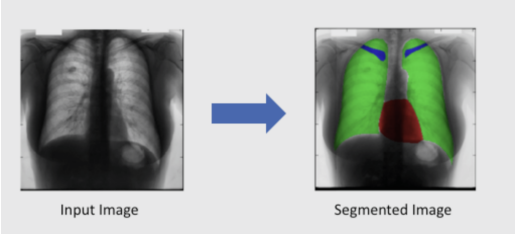
SGD is one of the oldest and most common optimization methods, others include Adam [KB14]. In Chapter 6 we will also discuss the Riemannian manifold alternative to gradient descent optimization.

2.4 Image Segmentation

With the progress of technology came the need and will to automate some aspects of medical image analysis. One of the most common ones, that will also be the topic of Chapter 5, is semantic segmentation. In semantic segmentation tasks we attempt to analyze an image on the pixel level attributing each pixel to a semantic category – like part of a dog or a cat. We illustrate examples of a semantic segmentation tasks in Figure 2.4.

Early methods of semantic segmentation operated directly in the pixel space where pixels were clustered according to their intensity values. Algorithms included *superpixels* where pixels were grouped with regards to their position and intensity levels. The resulting segmentation did not always coincide with human defined object classes but were a powerful initial tool for image processing. A different approach utilized active contours - colloquially referred to as "snakes". Under this paradigm, an energy function E_{snake} is minimized such that it creates

(a) Semantic Segmentation reproduced from [EVGW+]



(b) Medical example of semantic segmentation edited from [NLM+18]

Figure 2.4: Examples of Semantic Segmentation (a) Natural images from Pascal-VOC challenge; (b) Medical image example of lung segmentation

a connected graph representing the contour of the object depicted in the image [GGCV95]. Snakes, however can be computationally expensive and suffer greatly when not initialized ideally. GraphCut [GPS89] was another popular method that served as the base concept of numerous segmentation methods. Similarly to Active Contours, GraphCut tries to minimize an energy function depending on the color and coherence of the segmented structure.

Recent advances in machine learning allowed a paradigm shift in the ways we approach the task of semantic segmentation. Two methods are especially noteworthy, the Mask R-CNN [HGDG17] and the U-Net [RFB15] The Mask R-CNN belongs to the Region Proposal Network family. Initially a convolutional neural network is used to extract features from the input images. A region proposal network is then tasked to predict the anchor boxes location of the objects in the images. Mask R-CNN is characterized by a three-headed output: the first head outputs the predicted class of the object in the Region of Interest (ROI); the second offsets and adjusts the size of the bounding box to better match the object; while the the third produces a binary pixel level mask of where in the bounding box the object lies. The main contribution of the Mask R-CNN lies with the inclusion of the third output head that translates the ROI to an object mask.

In a quite different approach Ronneberger et al in their seminal U-Net paper [RFB15] assume a multi-level end-to-end convolutional architecture seen in Figure 2.5. Each image is passed through a multi level encoder-decoder network with skip connections between same level layers in the encoder and the decoder. The network is then tasked to predict a pixel-level segmentation mask and is trained with a categorical cross-entropy loss. The U-Net has been the de facto

default segmentation method for medical imaging tasks.

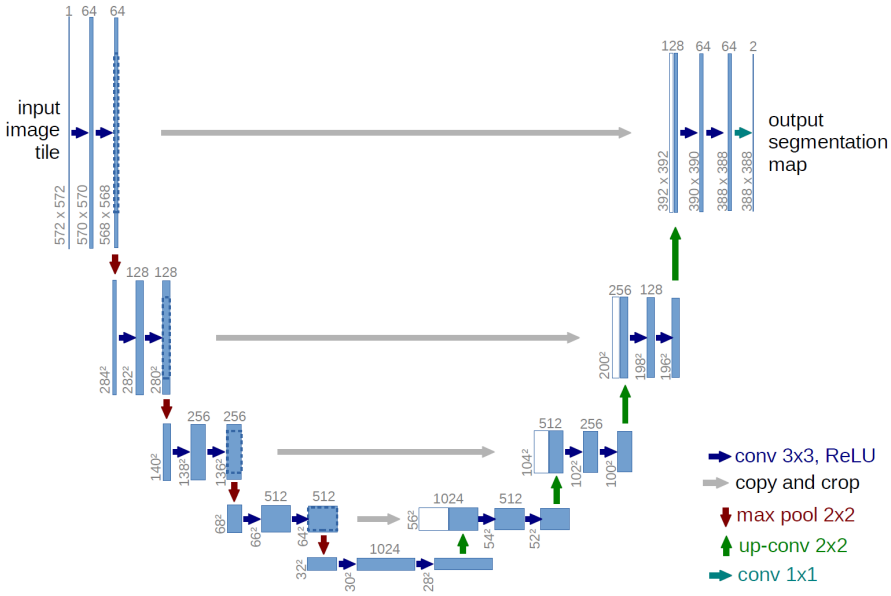


Figure 2.5: The U-Net Architecture. Image source[RFB15]

2.5 Hyperbolic Geometry & Manifolds

For an in-depth review of manifolds we invite the reader to refer to the textbook [Car97]. In this section we will briefly provide some key concepts of the used differential geometry appearing in Chapter 6.

Manifold: A manifold M of n dimensions is a generalization of the concept of the surface in a non-Euclidean space and possesses a curvature c at every point in spacetime. For our present purposes, we will consider only manifolds with constant curvature, that is, with the same value of c everywhere. The Lorentzian manifold group that this thesis considers is constantly flat ($c = 0$). We will also consider the Poincaré manifold, which is characterized by a negative curvature $c < 0$. Henceforth, we will be referring to the absolute value of c for ease of notation.

Tangent Space: The tangent space T_xM is a vector space that approximates the manifold M at a first order. All our models will lie on this Euclidean tangent space for ease of optimization.

Riemannian Metric: A Riemannian metric g is a collection of inner products $T_xM \times T_xM \to \mathbb{R}$. It can be used to define a global distance function as the greatest local bound of the length l of all the smooth curves γ connecting points $\mathbf{x}, \mathbf{y} \in \mathcal{M}$. Note that the length is defined as

$$l(\gamma) = \int_0^1 \sqrt{g_{\gamma(t)}[\gamma'(t), \gamma'(t)]}\, dt. \tag{2.7}$$

Geodesic: Geodesics are the generalizations of straight lines in Euclidean space and define the shortest path between two points of the manifold. They can also be defined as curves of constant speed.

Exponential map: The exponential map $exp_x : T_xM \to M$ around x defines the mapping of a small perturbation $v \in T_xM$ to a point in M s.t. $t \in [0, 1] \to exp_x(tv)$, which is the geodesic of x to $exp_x(v)$.

Poincaré Ball: Many works rely on a Poincaré ball, as do we in Chapter 6, as has been argued in [NK17, MLM+19, GBH18] that embedding the latent space on a Poincaré Ball – a hyperbolic space with negative curvature – allows to naturally embed continuous hierarchical relationships between data points. This follows from the qualitative properties of such a hyperbolic space:

1. The entirety of the Poincaré Ball $\mathcal{B}_c^d$ is contained within a hypersphere of radius $1/\sqrt{c}$ and dimensionality d, in what amounts to *compactification* of infinite space.

2. The distance function (and thus area element) of this space grows rapidly as one approaches the edges of $\mathcal{B}_c^d$, such that reaching the edge would require traversing an infinite distance in latent space.

3. This behaviour naturally emulates the properties of hierarchical trees, whose size grow exponentially as new branches "grow" from previously existing branches.

Quantitatively, the space $\mathcal{B}_c^d$ is endowed with a metric tensor g^c which relates to flat Euclidean space,

$$g^c(\mathbf{r}) = \left(\frac{2}{1 - c|\mathbf{r}|^2}\right)^2 g_e(\mathbf{r}), \tag{2.8}$$

where $\mathbf{r}$ is a d-dimensional vector in latent space and $g_e(\mathbf{r})$ the Euclidean metric. As a result, the distance element in $\mathcal{B}_c^d$ may be written, in spherical coordinates,

$$ds^2 = \left(\frac{2}{1 - cr^2}\right)^2 (dr^2 + r^2 d\Omega_d^2), \tag{2.9}$$

where $r = |\mathbf{r}|$ is the radius from the origin of the space and $d\Omega_d$ is the differential solid angle element in d dimensions. It easy to see that the distance element diverges as $r \to 1/\sqrt{c}$, thus encoding the infinite hypervolume contained near the edges of the Poincaré Ball. Furthermore, as $c \to 0$, the radius of the Poincaré Ball becomes infinity and $g_c(\mathbf{z}) \to g_e(\mathbf{z})$, up to a constant rescaling of the coordinates. Let $\gamma : t \to \gamma(t)$ be a curve in $\mathcal{B}_c^d$, where $t \in [0, 1]$ such that its length is defined by

$$L(\gamma(t)) = \int_0^1 \sqrt{ds^2(t)} dt = \int_0^1 \sqrt{\mathbf{v}^T(t)\, \hat{\mathbf{g}}^{\mathbf{c}}\, \mathbf{v}(t)}\, dt, \tag{2.10}$$

where $\hat{\mathbf{g}}^{\mathbf{c}}$ is the matrix form of g^c and $\mathbf{v} \equiv \frac{d\mathbf{r}}{dt}$ is the trajectory's velocity vector. In component form, this reads

$$= \int_0^1 \sqrt{\sum_{\mu=1}^{d} \sum_{\nu=1}^{d} \frac{dx^\mu(t)}{dt} g_{\mu\nu}^c \frac{dx^\nu(t)}{dt}\, dt} = \int_0^1 \left(\frac{2}{1 - cr^2(t)}\right)^2 \sqrt{\mathbf{v}^T \mathbf{v}}\, dt, \tag{2.11}$$

where x^μ represents each coordinate, $g_{\mu\nu}^c$ is the component form of $\hat{\mathbf{g}}^{\mathbf{c}}$ and in the last step we have used in Eq. (2.8). In hyperbolic space, "straight lines" are defined by *geodesics* $\gamma_g(t)$, *i.e.*, curves of constant speed and least distance between points $\mathbf{x}$ and $\mathbf{y}$. Thus,

$$\gamma_g(t) = \operatorname{argmin} \left[L(\gamma(t))\right]_{\gamma(0)=\mathbf{x}}^{\gamma(1)=\mathbf{y}} \quad \text{and} \quad \left|\frac{d\gamma(t)}{dt}\right| = 1. \tag{2.12}$$

With Eq. (2.11) and Eq. (2.12), one may show that the distance function $d^c(\mathbf{x}, \mathbf{y})$ between two points $\mathbf{x}$ and $\mathbf{y}$ on $\mathcal{B}_c^d$ can be computed to yield

$$d^c(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{c}} \text{arccosh}\left(1 + 2c\frac{|\mathbf{x} - \mathbf{y}|^2}{(1 - c|\mathbf{x}|^2)(1 - c|\mathbf{y}|^2)}\right). \tag{2.13}$$

Minkowski Space-Time (MST): Our main contributions are related to the Minkowski Space-Time (MST). We define the Minkowski Space-Time (MST) to be characterized by the metric of Equation (2.14) with the element -1 denoting the temporal dimension and $+1$ elements the spatial dimensions. MST constitutes a pseudo-Riemannian manifold as its metric is positive semi-definite as opposed to positive definite. Belonging to the Lorentzian group, the MST is a flat spacetime.

$$\eta_{\mu\nu} = \text{diag}(-1, +1, +1, +1). \tag{2.14}$$

2.6 Wrapped Normal

While embedding data on a Riemannian space with the use of a Riemannian VAE of Chapter 6, it is important to embed the used distribution in this space as well. Multiple ways have been proposed to perform this operation. We follow the wrapped normal distribution approach [GLA19, MLM+19].

For this, a normal distribution is mapped onto the manifold using the manifold's exponential map. Given a normal distribution $z_e \sim \mathcal{N}(0, \Sigma)$ and the Riemannian sample $z = exp_\mu^c(\frac{z_e}{\lambda_\mu^c})$, the distribution's density can be described as

$$\mathcal{N}_{\mathcal{B}_c^d}^W(z|\mu, \Sigma) = \frac{dv^W(z|\mu, \Sigma)}{d\mathcal{M}(z)} = \mathcal{N}(\lambda_\mu^c log_\mu(z)|0, \Sigma)(\frac{\sqrt{c}d_p^c(\mu, z)}{sinh(\sqrt{c}d_p^c(\mu, z))})^{d-1}. \tag{2.15}$$

With $c \to 0$ the Euclidean normal distribution can be obtained.

2.7 Reinforcement Learning

In Chapter 4 we develop a Reinforcement Learning methodology. Reinforcement Learning (RL) allows artificial agents to learn complex tasks by interacting with an environment E using a set of actions A. The agent learns to take an action a at every step (in a state s) towards the target solution guided by a reward signal r during training. The main goal is to maximize the expected rewards in order to find the optimal policy π^*. In Q-Learning, a state-action value function $Q(s, a)$ is used to approximate the value of taking an action in a given state. The Q-function is defined as the expected value of the accumulated discounted future rewards, which can be approximated iteratively as: $Q_{t+1}(s, a) = E[r + \gamma \max_a(Q_t(s', a'))]$. Here $\gamma \in [0, 1]$ is a discount factor that is used to incorporate the notion of uncertainty in future events. Mnih et al. [M$^+$15] proposed an approximation of the Q-function using a CNN by optimizing the network cost $L(\theta) = E\left[\left(r + \gamma \max_{a'} Q_{target}(s', a'; \theta^-) - Q_{net}(s, a; \theta)\right)^2\right]$. Q_{target} is a temporary fixed version of Q_{net}, which gets updated every N_{target} steps, used in order to avoid destabilization caused by rapid policy changes. While θ are the parameters of the network we optimize over.

In single-agent RL scenarios, individual models learn solely from states that result from the actions of an agent. Complementary to this, Multi-Agent RL (MARL) models learn from states that result from multiple agents dynamically interacting with their shared environment. In MARL models, there are K agents interacting with environment E. Each learns to take an action a_t^k during a state s_t^k using a reward signal r_t^k. Thus, the environment is subjected to the actions of all agents, as shown in Figure 4.2. Hence, the environment becomes non-stationary as action a_i in state s_k will not always lead to the same future, since the future state is also a function of the other agents. This causes a violation of the Markov assumptions needed for the formulation of a RL scenario as a Markov Decision Process (MDP). To address this issue, [FCAS$^+$18, RSdW$^+$18] proposed to establish communication between the agents, thus taking all agents' actions into account.

Any agent communication signifies the exchange of information or knowledge about the underlying Markov state of the environment. Communication between agents can be achieved explicitly via a communication protocol like in [FAdFW16], where a limited bandwidth chan-

nel is learned by the agents, or implicitly by sharing knowledge in the parameter space or by combining value functions [GEK17]. MARL scenarios can be classified as collaborative or competitive depending on the relation of the communication between agents. In this thesis, we define the collaborative scenario as agents that attempt to minimize a common loss function. Competition between agents signifies a scenario in which agents try to minimize their own loss function through increasing the loss function of other agents.

2.8 On the use of Causality

As we saw before, *medical imaging* is an umbrella term encompassing a number of imaging techniques including Magnetic Resonance Imaging (MRI), X-ray imaging, Computed Tomography (CT), and Ultrasound imaging (US), and is used primarily as supporting tool for diagnosis and monitoring of diseases. From a computational perspective, the community has engaged in a wide variety of tasks concerning the automated interpretation of medical imaging and enabling a range of applications. Machine learning (ML) shows significant successes for applications like automated localization and delineation of lesions and anatomies [RFB15, KLN$^+$17] as well as for the automated alignment of scans between patients and mapping of patient anatomy into canonical interpretation spaces [MV98, HKY20, GKGLF20, COL$^+$11]. Despite good in silico results, many of the approaches fail to translate into the clinical practice. While the reasons behind this can be complex and diverse, many have as a common factor the inability to adapt and be robust in clinical practice. Mapping this to the popular systems engineering framework of Technology Readiness Levels (TRL) [LGLV$^+$21a] as shown in Figure 2.6, medical imaging AI/ML applications often skip from TRL 4 – proof-of-concept– to TRL 7 –deployment– overlooking the very important TRLs 5 and 6 that make new systems robust to real world conditions. In Figure 2.7 we exhibit a finer grain view on the steps between TRL 6 and 7 that we deem to be important during the development of production-level medical imaging ML algorithms.

Concerning is the inability of commonly used AI/ML medical imaging methods to differentiate

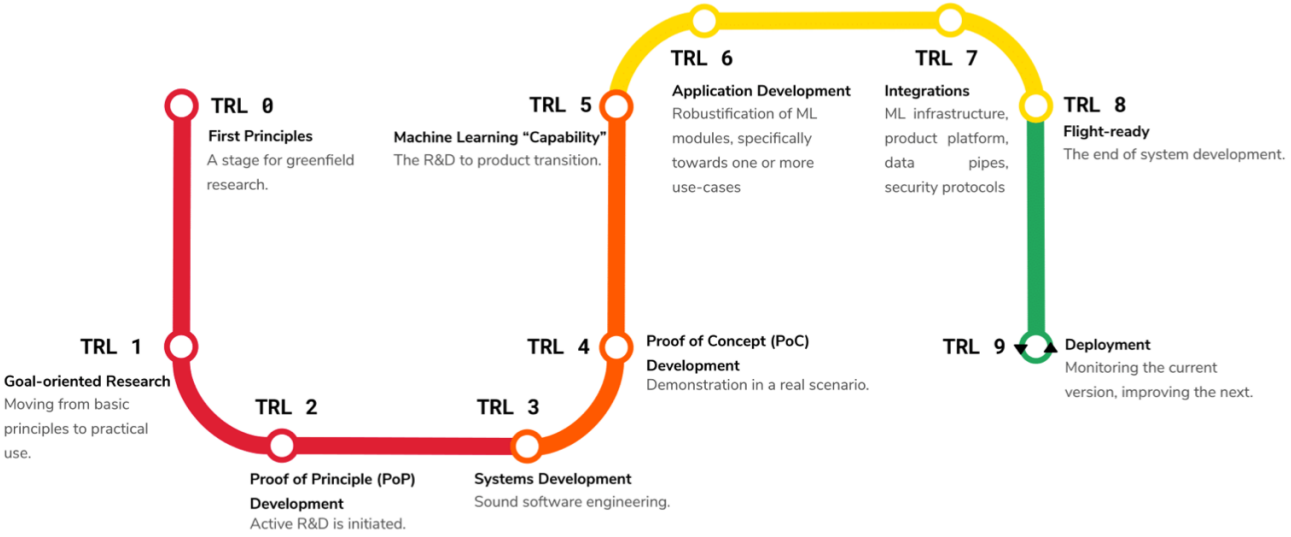


Figure 2.6: Technology Readiness Levels of ML Systems - most applications skip levels 5,6 where algorithms are made robust and production ready. Figure source: [LGLV+21a]

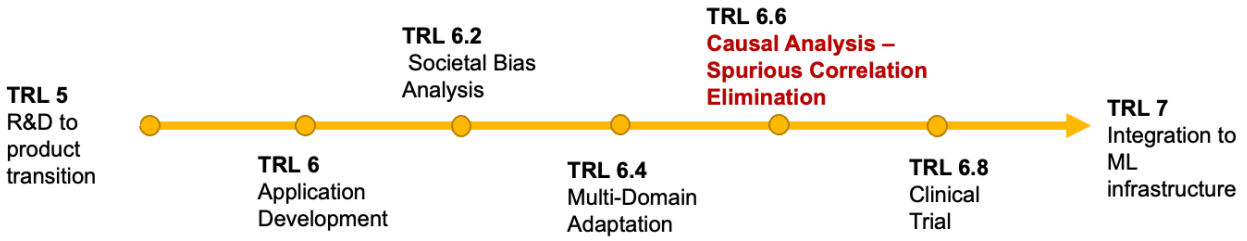


Figure 2.7: A in-depth view of the ML specific TRLs between TRL 6 and 7 in medical imaging - Inspired by [LGLV+21a]. In this chapter we focus on the TRL 6.6 and argue the need and benefits of causally robust ML algorithms

between correlations and causation, making potentially deadly mistakes in the process. For example, [DJL20] identified a number of approaches that claim to have been able to diagnose COVID-19 from chest X-rays but ultimately fail to do so as they were instead picking up spurious correlations like hospital identifiers and the ethnicity of the patient.

As pointed out by [CWG20], ML for medical imaging is susceptible to different domain shifts that affect algorithmic performance and robustness in new environments. Population shifts occur when the train and test populations exhibit different characteristics that might include the prevalence of diseases, for example prevalence of lung nodules in polluted urban environments is higher than in a rural setting. An ML model trained in one of the two settings is not able to condition on the true causal links that are unseen in the images, hence, treats both populations the same way.

Acquisition and annotation shifts affect the production of both images and ground truth labels used to train ML algorithms. For example the same MRI scanner used to acquire both images and annotations can lead to two different datasets, dependent on the scanner settings and medical beliefs of the radiologist performing the annotations.

Finally, data selection biases are especially prevalent in the medical domains where expanded datasets can be hard to create for ethical, economic, and legal reasons.

If acknowledged and mitigated by causal analysis, the aforementioned biases can help build robust and adaptable ML algorithms for medical imaging that minimize the chances of dangerous predictions due to spurious correlations. In essence, many of the negative phenomena seen in ML for medical imaging could be solved if the community expands its involvement with the TRL points shown in Figure 2.7, which suggest the inclusion of causal analysis as an important step. However, causal analysis is not commonly used for the development of ML applications for medical imaging. Thus, in the following sections we review and explore recent research in this direction and the use of causal analysis for medical image machine learning applications. In addition we will attempt to identify trends and lay out our beliefs for future directions of this field.

2.9 Mathematical Foundations of Causality

We first introduce some key concepts in causality required for the upcoming discussion of methods and Part II of this thesis. In Section 2.9.1 we discuss the concepts of Structural Causal Models and their parametrization as Directed Acyclical Graphs as introduced in computer science by J. Pearl. In Section 2.9.3 we define the notions behind Rubin's Potential Outcomes framework.

2.9.1 Structural causal models

Definition 2.1 (Structural Causal Model)**.** *A structural causal model (SCM) specifies a set of latent variables* $U = \{u_1, \ldots, u_n\}$ *distributed as* $P(U)$*, a set of observable variables* $= \{v_1, \ldots, v_m\}$*, a directed acyclic graph (DAG), called the* causal structure *of the model, whose nodes are the variables* $U \cup V$*, a collection of functions* $F = \{f_1, \ldots, f_n\}$*, such that* $v_i = f_i(PA_i, u_i)$*, for* $i = 1, \ldots, n$*, where* PA *denotes the parent observed nodes of an observed variable.*

The collection of functions F and the distribution over latent variables induces a distribution over observable variables: $P(V = v) := \sum_{\{u_i | f_i(PA_i, u_i) = v_i\}} P(u_i)$, In this manner, we can assign uncertainty over observable variables despite the fact that the underlying dynamics are deterministic. An example causal structure, represented as a directed acyclic graph (DAG), is depicted in Figure 2.8a.

Definition 2.2 (Submodel)**.** *Let* M *be a structural causal model,* X *a subset of observed variables with realisation* x*. A submodel* M_x *is the causal model with the same latent and observed variables as* M*, but with functions replaced with* $F_x = \{f_i \mid v_i \notin X\} \cup \{f'_j(PA_j, u_j) := x_j \mid v_j \in X\}$.

Definition 2.3 (do-operator)**.** *Let* M *be a structural causal model,* X *a set of observed variables. The effect of action* $do(X = x)$ *on* M *is given by the submodel* M_x*.*

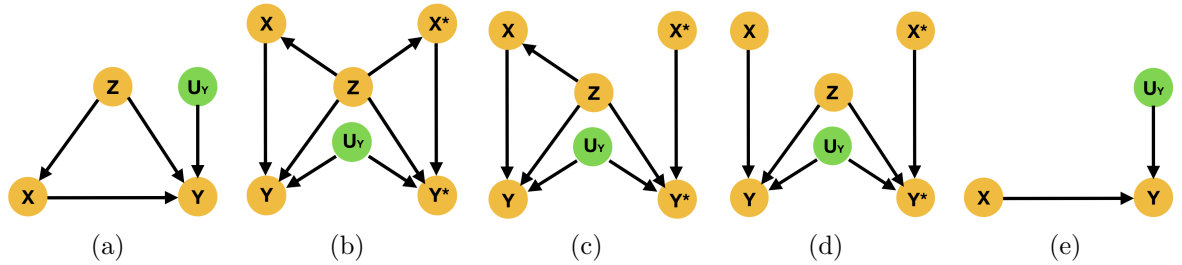


Figure 2.8: Orange nodes are observed, green latent. (a) Example SCM; (b) twin network of (a); (c) intervention in the twin network on node X^*; (d) interventions in the twin network on X & X^*; (e) Uncounfounded version of (a).

The *do*-operator, seen in Figures 2.8c and 2.8d, forces variables to take certain values, regardless of the original causal mechanism. Graphically, $do(X = x)$ means deleting edges incoming to X and setting $X = x$. Probabilities involving $do(x)$ are normal probabilities in submodel M_x: $P(Y = y \mid \text{do}(X = x)) = P_{M_x}(y)$.

2.9.2 Counterfactual inference

Definition 2.4 (Counterfactual)**.** *The counterfactual sentence "Y would be y (in situation $U = u$), had X been x", denoted $Y_x(u) = y$, corresponds to $Y = y$ in submodel M_x for $U = u$.*

The latent distribution $P(U)$ allows one to define probabilities of counterfactual queries, $P(Y_y = y) = \sum_{u|Y_x(u)=y} P(u)$. For $x \neq x'$ one can also define joint counterfactual probabilities, $P(Y_x = y, Y_{x'} = y') = \sum_{u|Y_x(u)=y,\ \&Y_{x'}(u)=y'} P(u)$. Moreover, one can define a counterfactual distribution given seemingly contradictory evidence. Given a set of observed evidence variables E, consider the probability $P(Y_x = y' \mid E = e)$. Despite the fact that this query may involve interventions that contradict the evidence, it is well-defined, as the intervention specifies a new submodel. Indeed, $P(Y_x = y' \mid E = e)$ is given by [Pea09] $\sum_u P(Y_x(u) = y')P(u \mid e)$. The following theorem provides an approach to computing such distributions.

Theorem 2.1 (Theorem 7.1.7 in [Pea09])**.** *Given SCM M with latent distribution $P(U)$ and evidence e, the conditional probability $P(Y_x \mid e)$ is evaluated as follows: 1)* ***Abduction:*** *Infer the posterior of the latent variables with evidence e to obtain $P(U \mid e)$, 2)* ***Action:*** *Apply $do(x)$*

to obtain submodel M_x, 3) ***Prediction:*** *Compute the probability of Y in the submodel M_x with $P(U \mid e)$.*

Definition 2.5 (Twin Model [BP94])**.** *Let $V = \{v_1, \ldots, v_n\}$ be observable nodes in the causal model and $V^* = \{v_1^*, \ldots, v_n^*\}$ the counterfactual duplicates of these. For every node v_i^* in the counterfactual model, its latent parent u_i^* is replaced with the original latent parent u_i in the factual such that the original latent variables are now a parent of two nodes, v_i and v_i^*. A visual representation of twin models is seen in Figure 2.8b*

Theorem 2.2 (Counterfactual Inference in a Twin Model [BP94])**.** *To compute a general counterfactual query $P(Y = y \mid E = e, do(X = x))$, one modifies the structure of the counterfactual network by dropping arrows from parents of X^* and setting them to value $X^* = x$. Then, in the twin network with this modified structure, one computes the following probability $P(Y^* = y \mid E = e, X^* = x)$ via standard inference techniques, where E are factual nodes.*

2.9.3 Potential outcomes – Average treatment effect

The potential outcomes framework introduced by [SNDS90] and [Rub78] explore the potential outcomes of a given intervention. Formally the framework is primarily interested in $Y(a), a \in \mathcal{C}$ where Y is the outcome given intervention a which belongs in the set of possible actions $\mathcal{C}$. For the simple case where we are interested in the outcome of a specific treatment, we can define use the following definitions:

Definition 2.6 (Treatment under Potential Outcomes)**.** *$\mathcal{D}_i$: Indicator of treatment intake for unit i*

$$\mathcal{D}_i = \begin{cases} 1, & \textit{if unit } i \textit{ received the treatment} \\ 0, & \textit{otherwise} \end{cases} \tag{2.16}$$

Definition 2.7 (Potential Outcomes)**.** *$\mathcal{Y}_{di}$: Potential outcomes for unit i depending if treatment has been applied or not*

$$\mathcal{Y}_{di} = \begin{cases} \mathcal{Y}_{1,i}, & \textit{Potential Outcome for unit } i \textit{ receiving the treatment} \\ \mathcal{Y}_{0,i}, & \textit{Potential Outcome for unit } i \textit{ not receiving the treatment} \end{cases} \tag{2.17}$$

Using these quantities we are able to define the the causal effect

Definition 2.8 (Causal Effect)**.** *Causal Effect of the treatment on the outcome for unit i is the difference between its two potential outcomes:*

$$\tau_i = \mathcal{Y}_{1,i} - \mathcal{Y}_{0,i} \tag{2.18}$$

In Equation (2.18) we define the causal treatment effect for an individual unit i. We can also define the average treatment effect that looks into a group of individual units $i \in \mathcal{G}$.

Definition 2.9 (Average Treatment Effect)**.** *τ_{ATE} is the difference between all treatment potential outcomes and all control potential outcomes*

$$\begin{aligned} \tau_{ATE} &= \frac{1}{N}\sum_{i}^{N} \mathcal{Y}_{1,i} - \frac{1}{N}\sum_{i}^{N} \mathcal{Y}_{0,i} \\ &= E[\mathcal{Y}_{1,i} - \mathcal{Y}_{0,i}] \\ &= E[\tau_i] \end{aligned} \tag{2.19}$$

In literature we can find many variations of this measure like the individual treatment effect (eg. [MLP21]) where we look on the treatment effect on an individual rather than aggregate over a population. In the interest of conciseness we will not be be exploring the full range of possible variations upon the ATE in this section but just acknowledge that they exist and call upon them depending on the method we are discussing.

One last concept required for our discussion is *propensity score*. Introduced by [RR83] it is defined to be the probability of treatment assignment conditioned on observed covariates. Mathematically it can be described as $e = P(T \mid X)$ where X are the covariates and T the

treatment. Commonly used as a matching criterion to form sets of treated and untreated subjects that are close in the covariate space; these sets are subsequently used to estimate the causal effect of the treatment.

2.10 Causal discovery in medical imaging

Causal discovery is an open and challenging research problem, directly touching the most fundamental aspects of scientific exploration; the discovery of causal relations [VCB21]. In this setting we are trying to estimate the mechanisms that describe the causal links between variables from data. In order to make the problem of causal discovery tractable, common causal discovery methods make a series of assumptions that can be summarized as:

- Acyclicity - we are able to describe the causal structure as a Directed Acyclical Graph (DAG)
- Markovian - all nodes are independent of their non-descendant when conditioned on their parents
- Faithfulness - all conditional independences are represented in the DAG
- Sufficiency - any pair of nodes in the DAG have no common external causes

Moreover, the vast majority of approaches tackling causal discovery formulate the problem as graph modeling challenge. We refer the reader to [GZS19, VCB21] for a more thorough review of causal discovery methods based on graphical models. For the purposes of this section we note that methods are often categorized based on their approach into *constraint-based*, *score-based* and *optimization-based* [ZARX18] methods.

Constraint-based methods relate to approaches like [SGSH00] PC and Fast Causal Inference (FCI) that test conditional independence between factors to assess their causal links. The main intuition is that two statistically independent variables are not causally linked. First, pairwise independence is evaluate to determine the undirected structure. Following this conditional

independence is tested to orient the links between the nodes. If two nodes fail this test, then they can be added to the separation set of each other used to orient colliders - nodes of the causal graph with two or more incoming links. The main difference between the aforementioned PC and FCI algorithms is FCI's ability to be asymptotically correct in the presence of confounders. Both of them are however limited to the causal *equivalence classes*, *i.e.*, causal structures that satisfy the same conditional Independence. A method that searches over the space of possible equivalence classes is the Greedy Equivalence Search (GES) [Chi03] officially considered a score-based method which uses the Bayesian Information Criterion (BIC). Similarly, [PSD+20] use a score based approach that characterizes the acyclicity constraint of Directed Acyclical Graphs as a smooth equality constraint.

In the computer vision field, visual causal discovery has been spearheaded by tasks related to the CLEVRER dataset [YGL+20] where ML algorithms are asked to understand a video scene and answer counterfactual questions.[LTA+20] learn to predict causal links from videos, by parameterizing the causal links as strings and springs where the algorithm predicts their parameters. [LMZW22] perform a similar task by inferring stationary causal graphs from videos. Furthermore, [NBS19] develop a causal discovery method based upon the use of attention-enabled convolutional neural networks. [KCW+22] address the task of causal discovery by training a neural network to induce causal dependencies by predicting causal adjacency matrices between variables. Finally, works like [GGLT21] analyze how humans perform the task of causal discovery in relation to how machine learning approaches the same task.

In the sub-field of medical imaging, causal discovery has not been explored to its fullest potential. Most works involve functional Magnetic Resonance Imaging (f-MRI). f-MRI is able to highlight active areas of the brain, and as during different physiological processes (working or resting) activate a series of brain areas. Causal discovery in f-MRI attempts to identify causal links among neural processes. [SRRZ+18, SRRZG19] diverge from the usual DAG paradigm and introduce the Fast Adjacency Skewness (FASK) method which exploits non-Gaussian features of blood-oxygen-level-dependent (BOLD) signals to identify causal feedback mechanisms. In a very interesting work, [HZZ+20] parametrize both causal discovery in f-MRI and domain adaptation as a non-stationary causal task that they then proceed to solve. In the field of

medicine but not imaging [MC00] extract causal relations from medical textual data.

[BUvM+17] looked into applications on f-MRI and extracting causal relations between neural processes by applying the most common causal inference techniques including dynamic causal modeling, Bayesian networks, transfer entropy and ranked them based on their perceived suitability and performance on the tasks at hand. Similarly, [CBTRDE21] explore deep learning methods for schizophrenia analysis and outcome prediction, mostly with the use of f-MRI inputs, and argue for the need of causally enabled ML methods to produce plausible hypothesis explaining observed phenomena. Recently [CRBC22] propose deep stacking networks (DSNs), with adaptive convolutional kernels (ACKs) as component parts to aid the identification of non-linear causal relationships.

Furthermore, [JLH+18] discuss bivariate causal discovery for imaging data. The authors develop a non-linear additive noise model for that they show a causal discovery task with genetic data and for causal inference in the Alzheimer's Disease Neuroimaging Initiative (ADNI) MRI data. [RCP21] propose a Structural causal model that encodes causal functional relationships between demographics and disease covariates with MRI images of the brain in an attempt to identify Multiple Sclerosis (MS).

While the field of medicine is governed by causal relations between physiological processes, little work has been done in the field of medical imaging and causal discovery. We believe that this is mostly due to the significant inherent difficulty of the task at hand, in conjunction with the lack of the required meta-data that are needed to characterize causal links that are not visible in the image. In other words, we do not think that images by themselves possess enough information for identification of causal links but we are adamant that they can serve as a useful tool and source of extra information when combined with medical knowledge. It is, however, important to note the potential the aforementioned time-series based causal discovery methods exhibit as f-MRIs can themselves be thought of as time-modulated events, and be used in conjunction with other time-based modalities.

Finally, relating back to Figure 2.7, causal discovery can aid bring to light causal links that were not previously known. It can further help, probe the beliefs of other algorithms and hence

root out spurious correlations or the encoding of societal biases that their creators carried. As such we believe that causal discovery can assist the completion of TRLs 6.6 through audits of the practitioners beliefs. Most importantly, however, in the subsequent sections we will be discussing causal inference, that assumes the existence of high quality causal diagrams. Causal discovery is responsible for establishing these diagrams and hence indirectly contributing significantly at TRLs 6-6.4.

2.11 Causal Inference in medical imaging

While causal discovery using medical images is limited to some applications involving f-MRI scans, causal inference is significantly more active as an area of research. Highlighting its importance, [CWG20] argue that causal inference can be used to alleviate some of the most prominent problems in medical imaging. They argue that acquisition and annotation of medical images can exhibit bias from the annotators and curators of the datasets. As such, causality aware methods can learn to account for such biases and reduce their effects. Moreover, as the training datasets represent a limited population with specific characteristics, medical imaging algorithms are susceptible to population, selection and prevalence biases if not properly controlled for these variations. These biases could for example arise when an algorithm is trained by a vast majority on data that come from a given geographical region X, then it implicitly learns the prevalence of diseases for that group; if it is deployed in a different region with a population that is characterized by different genotype and phenotype characteristics, the biases that the algorithm has learned could lead to mis- or under- diagnosis of diseases.

In our exploration of the use of causal inference in medical imaging literature we identified five main sub-fields of research that leverage causal insights. We found that causal inference is overwhelmingly used to contribute to fairness, safety and explainability of the existing approaches. There are some albeit limited uses in generative modeling of medical images, domain generalization and out-of-distribution detection. As we will expand in the following sections, we believe these areas are ready for more applications of causal inference.

We note that all the works mentioned below have causality as a key part of their proposed methodologies. We extended our review not only to peer reviewed publications but also to notable preprints that appear to have produced significant discussion in the machine learning medical imaging community.

2.11.1 Medical Analysis

We continue our literature review with causally enabled methods for medical analysis that utilize imaging data.

In [WCD21] the authors use a normalizing flow-based causal model similar to [PCG20] in order to harmonize heterogeneous medical data. Applied to T1 brain MRI for the classification of Alzheimer's disease, the method abides by the abduction-action-prediction paradigm to infer counterfactuals which are then used to harmonize the medical data. [PW21] circumvent the identifiability condition that all confounders have to be known and measured by leveraging the dependencies between causes in order to determine substitute confounders; they apply their method in brain neuroimaging for Alzheimer's disease detection. On a similar note, [ZDT+21] issue an alternative to expectation maximization(EM) for dynamic causal modeling in f-MRI brain scans. Their approach is based upon the multiple-shooting method to estimate the parameters of ordinary differential equations (ODEs) under noisy observations required for brain causal modeling. The authors suggest an augmentation of the aforementioned method called multiple-shooting adjoint method by using the adjoint method to calculate the loss and gradients of their model.

[CFL+22] propose a neural score matching method for high dimensional data that could be potentially very helpful for the development of medical imaging applications as they develop methods for causal inference in high dimensional settings; allowing thus the use of medical images in a more straightforward way that avoids pre-processing them to a lower dimensional latent space. Finally, [RHH+10] identified six problems that causal inference can assist in solving in the field of functional MRI analysis.

2.11.2 Fairness, Safety & Explainability

Another application of causal inference in the field of medical imaging revolves around fairness, safety and explainability, directly related to TRLs 6.2 6.6 and 6.8. Medical tools like medical imaging analysis have a significant impact on the well-being of people affected by their use. Doctors and patients alike need to be able to trust the AI/ML methods in order to use them, while contrary to other AI/ML applications unwanted bias and poor performance can often be deadly. As such, the need to have fair, safe and explainable algorithms arises. Causal inference is a great tool to analyze black box AI/ML methods and make sure that they are not carrying unwanted societal biases and mitigate any robustness problems that might arise.

In this field [KSMZ+20] accompany their ML algorithm to detect polyps in human intestines with a causality inspired analysis on the effectiveness of their method. Similarly [dSGS+20] use generative model produced brain MRI images of brain atrophy to evaluate and explore different causal hypothesis on brain growth and atrophy. On a similar note [LSR+21] develop a causal inference-based method to search associations in genomics data from the UK Biobank. Additionally, [BPG+21] employ causal analysis to explain the performance of their brain structure segmentation network. Similarly [SWTB21] employ mediation analysis to identify the units and parameters of radiological reports that influence their classifier's outcomes; this method is applied on chest X-rays. In their work [ZDL+21] develop a Bayesian causal model to interpret the outputs and functionality of their Faster-RCNN based pelvic fracture classifier in CT images. Their Bayesian causal model matches lower confidence predictions with higher confidence ones and then updates the prediction set based on these matching. [GSCHH] explore the effect of uncontrolled confounders in medical imaging applications and observe that regardless of task and architecture, total confounding can be used to explain the difference in performance between development of the models and real life applications. [AMAK22] evaluate new metrics to quantify the effect of spurious correlations in age regression from hand X-rays. They show that only under certain conditions these metrics can be trusted and call for a paradigm shift in the effort to identify spurious correlations

Concerning fairness, [CCL+21] discuss the effects of biases in medical ML and how biases like

image acquisition, genetic variation, and intra-observer labeling result in healthcare disparities. They go on to argue that causal analysis in medical ML can greatly help mitigate such biases. Expanding on this argumentation, [HDES+22] argue that information fusion is key to achieve greater transparency and safety in medical imaging ML applications. In a slightly different position paper, [SCVH21] argue for the standardization of medical metadata in order to assist causal inference techniques in biomedical ML. Along similar lines, [GML21] argues for the use of causal intuition when designing medical imaging datasets. Meanwhile, [VRRK22] uses causal inference to estimate the necessity and sufficiency of the type and quantity of data to include in a medical imaging dataset in order to improve model performance under strict computational and financial constraints. In addition, [VSGS21] contend that causal analysis can help alleviate biases and provide the necessary trust to medical imaging application in deep space manned missions.

Finally, [BJG22] investigated questions of algorithmic fairness in medical imaging ML under a causal prism, focusing on the issue of under-diagnosis they highlighted some issues that warrant more attention in prior pieces of literature. [SHK+22] perform a thorough evaluation of the biases and unfairness that can arise in cross-hospital deployment of medical ML solutions asserting a causal analysis as a potential method to alleviate these issues. [BPP+21] use propensity scores (see section 2.9.3) to quantify diversity due to major sources of population stratification and hence assess fairness.

2.11.3 Generative methods

Generative modeling attempts to learn variable interdependencies such that the model is able to generate realistic samples that abide by certain characteristics aiding the admittance past TRLs 6.4 and 6.8. Variational Autoencoders [KW13], Generative adversarial networks [GPAM+14] and normalizing flows [DSDB16] are examples for approaches that try to estimate the underlying data distribution from which they then sample to produce new data. Causal inference in generative modeling is a relatively underdeveloped field, especially in the context of medical imaging due to the inherit difficulty to acquire good quality training signals for the counterfac-

tual samples.

[GVMB22] develop a two stage methodology where Tuberculosis infected lung CT images are analyzed in a disentangled manner and produce counterfactual images depicting how the patient would look like if they were healthy. Contrary to other approaches the authors use a DAG to represent the image generation process and parametrize it using a neural network such that sampling and use of it is straightforward for the counterfactual generation step. [PCG20] developed a normalizing flow model to perform the abduction step in an abduction-action-prediction counterfactual inference task and are able to generate plausible brain MRI volumes. Reynaud et al. in [RVD+22] assume a different approach and develop a generative model based on Deep Twin Networks [VKGL21]. Performing counterfactual inference in the latent space embeddings, the authors are able to generate realistic Ultrasound Videos with different Left Ventricle Ejection Fractions. Their approach is similar to [KSDV18] in the sense that a GAN is used to provide a training signal for the generated, counterfactual samples. Moreover, [KHN+22] in a methodologically similar note, generate counterfactual images to guide the discovery of medical biomarkers in brain MRI volumes.

Finally [SKL+22] use deep diffusion model to ask counterfactual questions and generate hard to obtain medical scans. These, in turn, are used to augment existing datasets for other downstream tasks.

2.11.4 Domain Generalization

One of the most promising areas where causal reasoning can be applied in the field of medical imaging is Domain Adaptation and Out-of-Distribution detection, directly associated with TRL 6.4. If we model the generative process that results in a medical image and include factors like the medical history, the disease, imaging domain, etc. we can then go on and interpret domain generalization and adaptation as a model that is able to perform well under different treatments in the imaging domain parameter, as argued by [HZZ+20]. In their paper Huang et al model domain adaptation as a non stationary change in the underlying causal graph and propose methods to identify and resolve these changes. [ZDSK+21] analyze the domain shifts

experienced in clinical deployment of AIML algorithms from a causal perspective and then proceed to investigate and benchmark eight popular methods of domain generalization. They find that domain generalization methods fail to provide any improvement in performance over empirical risk minimization in situations where we find sampling bias. Similarly [FPKK22] model the causal relationships leading to the medical images and create synthetic datasets in order to evaluate the transportability of methods to external settings where interventions on factors like ages, sex and medical metrics have been performed.

[OCL+21] apply a causal analysis on the problem of domain generalization in segmentation of medical images. They first simulate shifted domain images via a randomly weighted shallow network; then they intervene upon the images such that spurious correlations are removed and finally train their segmentation model while enforcing a domain invariance condition. [VLT21] develop a method to reuse adversarial mask discriminators for test-time training to combat distribution shifts in medical image segmentation tasks. In their discussion of their method they explain the good performance of their method under a causal lens. Finally [KØP+19] build a causal Bayesian prior to aide MRI tissue segmentation to generalize across different medical centers.

2.11.5 Out of Distribution Robustness & Detection

We ought to consider the use of causal reasoning and inference on out of distribution detection tasks aiding TRLs 6.6 and 6.8. Commonly seen as anomaly detection tasks many methodologies attempt to learn the underlying distribution of "normal"- *in-distribution* data and assess whether test *out-of-distribution* "pathological" samples belong to the same distribution or not. In a related task, researchers sometimes treat the *out-of-distribution* as a robustness criterion and attempt to develop methodologies that can operate equally well in all domains. It is evident that an alternative, where we learn the causal dependencies that make samples that are considered *out-of-distribution* can yield significant benefits in the field of anomaly detection. In this sub-field we only found a few works exploring this approach in medical imaging but we believe that more works will appear once the community gets more familiar with the benefits

of causal reasoning.

[YLH+21] give a causal explanation to diversity and correlation shifts and proceed to benchmark out-of-distribution methods, showing that the aforementioned shifts are the main components of the distribution shifts found in OOD datasets. Similarly [YXLL21], introduce an influence function and a novel metric to evaluate OOD while analyzing their contributions from a causal standpoint. [LSW+21] propose a causal semantic generative model in order to address OOD prediction from a single training domain. Utilizing the causal invariance principle they disentangle the semantic causes of prediction and other variation factors achieving impressive results.

On the robustness of medical procedures [DZK+22], built a causal tool segmentation model that iteratively aligns tool masks with observations. Unable to deal with occlusions and without leveraging temporal information the authors of this recent work also comment on the future next steps of robust causal machine learning tools.

Even though not directly applied in medical imaging, causality has been playing an important role in the algorithmic robustness literature as seen in [ZGL+22, PSWB18]. Meanwhile the machine learning for medical imaging community has shown great interest in developing robust algorithms, [HMT21, HWB+19]. We hope that these two communities will soon come together and use causal reasoning in making machine learning for medical imaging algorithms robust.

Chapter 3

A Case Study

In this chapter we discuss a case study to give prominence to the need of an Autonomous Diagnostic System (ADS). We want to aid the motivation and understanding of the practical applications of this thesis. The chapter below is based upon [VSGS21] where the author was the primary driver of the case study setting out and discussing the potential of automated diagnostic systems in deep space exploration scenarios.

In this case study we point out that future short or long-term space missions require a new generation of monitoring and diagnostic systems due to communication impasses as well as limitations in specialized crew and equipment. Machine learning supported diagnostic systems present a viable solution for medical and technical applications. We discuss challenges and applicability of such systems in light of upcoming missions and outline an example use case for a next-generation medical diagnostic system for future space operations. Additionally, we present approach recommendations and constraints for the successful generation and use of machine learning models aboard a spacecraft.

3.1 Introduction

Space agencies have a renewed drive to take human space exploration beyond Lower Earth Orbit (LEO) and into deep space. NASA's Artemis program outlines a clear path to return to

the Moon and to go beyond to Mars [NAS17]. Additionally, recent successes in the commercial space sector by major players such as SpaceX and Blue Origin make human spaceflight more accessible, affordable and future long term-missions a reality. Yet future long duration spaceflight require systems that are independent of LEO operations such as constant communication, the ability to transfer large amounts of data via multiple systems in a relatively short time-frame or the ability to request and exchange crew if needed. On Earth, ML and machine automation is already driving the next industrial revolution and resulted in fully autonomous industrial processes in domains such as agriculture as well as manufacturing [AAUS+19, YKBT19]. Spaceflight itself, however, is far behind such advances. Here we discuss challenges ML supported systems face in the space domain as well as the applicability and advantages of ML systems on a spacecraft. We highlight aforementioned items via an example of an autonomous medical system and describe an infrastructure for the successful development of such systems.

3.2 Challenges for Machine Learning aboard Spacecraft

Space is challenging and manned space exploration is dangerous and unforgiving. Moreover, the long term effects of human presence in micro gravity are still not fully known. Current space missions do rarely include a medical officer among the crew and rely on specialists who are also trained in emergency medicine. While this is sufficient for minor and trained emergency cases, it does not allow for more serious and complex medical treatments. Health related checks and emergencies are handled in a telemedicine regime where instructions are communicated to the astronauts via ground to spacecraft channels. In extreme emergencies the astronauts can always make the journey back on Earth. As missions veer further away from Earth, returning for medical treatment and relying on simultaneous communications becomes infeasible as both distance and communication latency increases significantly. For a deployed ML system aboard a spacecraft there are several challenges to consider: a) limited live testing abilities available, and testing in the form of payloads on missions are expensive; b) systems that are deployed need to be at a high Technology Readiness Level (TRL) [LGLV+21b]; c) environmental effects may influence deployed systems. For example, how ionizing radiation can affect deployed space

capable hardware [PB05], affecting the consistency of sensor behavior such as early or late sensor fusion; leading to potentially corrupted data to be processed e) payloads are constrained by weight, so shipping large compute infrastructures is infeasible; f) fully autonomous applications can be found mainly on controlled environments that assume almost complete access to information and environmental parameters; g) and more importantly the lack of labeled data for each task along with limited interpretability and explainablity of current ML systems add to the complexity. Contrary to that, any space-faring vehicle is faced with extreme environmental conditions that not only are hard to control or predict, but in some cases are challenging to human scientific comprehension. Hence, ML systems must be as robust as possible to changes of their operating environments.

3.3 Applicability & Advantages

Incorporating autonomous processes in a spacecraft is a complex task as technology and the associated needs constantly develop. Some key points, however, are: (1) Reduce Latency and Earth Reliance - A significant amount of operations aboard modern day spacecrafts and the ISS, require the constant communication with mission control on Earth [Dem17]. As space exploration expands further away from LEO and the Moon, the delay in communication represent an insurmountable obstacle for remote guidance, control and communications. Characteristically, we note that the round-trip time for current communications with Mars ranges between 5 and 20 minutes depending on the state of the two planets orbits [NAS20]. ML, thus, can resolve the dependency on communications and perform the mission critical information processing aboard. (2) Adapting Maintenance - Modern spacecrafts both manned and unmanned, constitute extremely complex systems that even with the use of automated checks are still prone to faults, especially when faced with extraordinary circumstances. The crew might not be able to solve the issues by themselves. To tackle this we believe that ML systems can perform functions that transcend anomaly detection and fault prediction. An ideal deployed ML solution needs to account for automated maintenance and resolution of faults both in hardware and in the software of the spacecraft. (3) Reduce latency for functions that don't require manual

intervention and can be conducted at scale. For example, automated checksums for data or models (4) Recent advances in ML like bit quantization, pruning, and hardware approximations [WDZ+19] enable inference on resource constraint edge devices which can also be similar to the target hardware in space [HJL+20].

3.4 Medical Use Case

In the previous sections we have set out some challenges as well as advantages for the use of ML on spacecraft. In this section we will be exploring the medical use case. As mentioned before, medical treatment in space relies on communication with the earth, a reliance we need to remove as we progress into deep space missions. A major part of medical treatment is the ability to inspect bodily functions in a non invasive manner through the use of medical imaging devices. MRI, CT and X-Rays are widely used on earth but are not ideal for space use, as we explain below. To this effect, we propose the use of Point Of Care Ultrasound (POCUS) as a lightweight non ionizing imaging solution.

ML Enabled Imaging: Medical imaging devices are often slow, requiring significant resources to operate and store and depend on the use of ionizing radiation (e.g. X-Rays/CT Scans) that has negative effects on patients and doctors alike [FDA21]. One notable exception to the above constraints is Point Of Care Ultrasound (POCUS). Mobile device enabled Ultrasound (US) probes are safe for patients and doctors, require minimal resources to operate —weighing less than 1kg— and offer real time imaging capabilities [GEH21]. However, acquiring medically significant images with a POCUS probe is non trivial and requires expert operators.

As expert users might not always be available in deep space missions we envision the use of ML enabled POCUS probes that guide the user to medically relevant areas.

Approaching the task of navigation one can identify straight away the need to be able to determine where the intelligent agent is with respect to the patient's body. Basic anatomy knowledge by both patient and other crew members is assumed for the purposes of this application. Navigating to points of medical interest, though is non trivial. Following medical practice of first

identifying standard planes and anatomical landmarks; RL solutions of disembodied agents have been proposed in [AOL+19b, VAK+19] to perform the above tasks, providing high accuracy and low computational constraints. [MBS19, HAT+20, LWX+21] expand the RL methods including the degrees of freedom of the US probe into the agent's action space and optimize directly on identifying landmarks while controlling a virtual probe. All the above methods find themselves limited to the anatomies that they have been trained on, with their computational burden rising significantly when trained to find standard planes or landmarks of multiple anatomies.

As such we believe that a hierarchical approach to the problem would help keep complexity and computation low enough for space craft applications while maintaining high enough performance. Approaches like MAX-Q [Die00] have withstood the test of time and have theoretical guarantees on convergence and the ability of the algorithm that given a set of subtasks it can find the global optimal policy. In short, MAX-Q constructs a hierarchical action tree that a higher level agent uses to sets out subtasks for other agents. In this fashion one can design a fully autonomous agent that collects information in form of images through POCUS navigation, assesses biomarkers based on measurements derived from landmarks, and regresses the diagnosis.

Furthermore, [EJRB+18, DG16, RG18] have shown that incorporating modules that aid agents "imagine" how scenes would look like from different points of view, increases their performance in scene understanding. Hence, we are firm believers that incorporating such modules in approaches like [MBS19] would increase their performance capabilities.

Finally, as medical applications are of critical importance, appropriate checks and balances should be put in place to avoid any harm to the crew. We believe that a rule based set of parameters should be developed in collaboration with physicians that would constitute a fallback system. Having no dependence on learned or inferred information we are able to guarantee a basic level of care to the astronauts in case all other systems fail. On top of the safety-net rule based system, a causal inference infrastructure can be used to assess the the probabilities of causality (Sufficiency and Necessity) [BWZ+18, KLRS17, LSM+17, VRRK20,

BBH+21]. Causal counterfactual inference enable us to assess the causal links and potential outcomes of treatments providing, thus, a more informed decision process.

3.5 Infrastructure considerations for Machine Learning systems on Spacecraft

Developing novel ML methods to provide mission critical treatments to astronauts is a hard task. In the previous section we set out some thoughts on a potential application for medical imaging aboard a spacecraft. In this section we will be briefly exploring infrastructure considerations that are directly related to the operation and development of medical imaging ML solutions. These considerations are also apparent on earth bound ML applications but gain increased importance due to the edge cases that they are called to operate on during spaceflight.

Data Collection Medical data acquisition is faced with a series of challenges. The privacy of the patients donating their data has to be protected throughout the data acquisition and model development processes. In order to comply with all legal obligations, we propose the data to have NIHR [NIH21] approved methods in place for making patient data anonymous. Meanwhile, storage of data must be in a HIPAA [HIP21] compliant platform. Federated learning also present promising methods in regards to secure data processing [KMY+16]. Going beyond the need of anonymization and maintenance we would like to draw the reader's attention to two constraints of increased importance on medical ML applications

Domain Shift: Domain shift robustness is an open topic of ML research that focuses on making existing methods robust to distributional shifts from the underlying training data domain [LZWJ16, ZWW+20]. In medical applications and more prominently in medical imaging applications domain shifts are easier to manifest and harder to overcome. The first major factor in this phenomenon comes from the equipment used. Medical equipment, when installed in hospitals and health centers is configured by the seller to the exact specifications of the attending physicians, as such, two different doctors using the same base equipment on the same

patient can result in two quite different images. The standardization of medical equipment configurations in space missions would aid to decrease the equipment-induced domain shifts.

Another source for domain shift comes from the training patient characteristics. Different populations exhibit different medical characteristics. Hereditary traits as well as phenotype derived attributes place an individual in different medical risk groups and force pathologies to manifest with different probabilities. These differences are often picked up by ML algorithms as unwanted inductive biases, skewing the learned conclusions. In the context of space missions, astronauts have diverse backgrounds both genetically and in terms of phenotype, as such models trained on a general population not representative of the characteristics of the crew can provide skewed estimations of biomarkers vital for diagnosis. It is imperative, then, that the deployed models be calibrated to the genetic background and phenotype of the astronauts.

Anomalous phenomena: As stated in the above motivation, the effects on human well-being stemming from prolonged exposure to cosmic radiation and other space related phenomena are not fully understood. It is unknown, hence, how the human body might react to adverse conditions. ML applications cannot, then, be expected to cover the full range of scenarios, on the contrary they should be expected to fail when presented with data that constitute an anomalous effect. In order to combat potential failure cases of the medical systems on-board a spacecraft, we suggest accompanying any ML algorithm with an anomaly detection mechanism (perhaps probabilistic or out-of-distribution mechanics), that flags non standard bio markers [WBR+20], and an active learning [Set12] feedback pipeline such that new, medical phenomena are incorporated in the evolution of the medical algorithms on-board.

Computational Considerations High communication latency creates the demand for on-board computation. However, cosmic radiation and weight offers strict constraints on the available on-board compute resources. Cosmic radiation is able to produce errors in modern scale computing units, while weight restrictions exist in all parts of a mission, from lift-off to grounding. ML applications should, then, be able to function without heavy computational needs, a non-trivial feat especially on medical imaging algorithms. This reinforces our case towards the use of POCUS as they have been shown to be computationally and physically

lightweight. Recently [HJL+20] has shown that NVIDIA's Jetson Xavier modules are able to withstand a significant level of proton-based radiation, making them optimal candidates for on-board inference and fine-tuning infrastructure. NASA has also awarded an SBIR contract to Numem [Gha20] to develop a radiation hardened DNN co-processor for a wide variety of ML applications, from low power machine vision to healthcare.

We note that of equal importance are storage capabilities that unfortunately are also compromised by weight and radiation restrictions when used in orbit. While storing models pre-trained on earth is a viable short term solutions, we foresee the need for fine tuning and inference on newly collected data. As such we focused our case study on the computational capabilities rather than storage.

3.6 Summary

ML-supported medical diagnostic systems on spacecraft are necessary for long-term space mission to overcome limitations in ground-to-spacecraft communication and lack of qualified medical crew. In this case study, we outlined the importance of incorporating ML-enabled medical applications on spacecrafts and considered challenges that need to be overcome in terms of anomaly detection and domain adaptation. Finally, we discussed ideas on how to augment existing POCUS algorithms such that they constitute a complete diagnostic system. We gave the example of a hierarchical RL method that also includes causal inference checks and balances such that the health and safety of crew members is guaranteed. Having set out the motivation of our applications we now turn towards developing the components outlined in Figure 1.1.

Part I

Exploration

Chapter 4

Exploring & Finding

As set out in the Introduction of this Thesis and Figure 1.1 we will first tackle the problem of exploring our environment and finding points of medical interest. Of course, this is not a trivial task, we will be suggesting a novel reinforcement learning approach that allows multiple agents to communicate implicitly with each other and aid each other in achieving the end tasks. For the benefit of the reader Figure 4.1 shows the relative position of this chapter in our efforts towards building a ADS. This chapter is based upon [VAK+19] published in MICCAI 2019 where the author conceived of the study, developed the appropriate tools and methods, ran the experiments and led the writing of the paper

The exact localization of anatomical landmarks in medical images is a crucial requirement for many clinical applications such as image registration and segmentation as well as computer-aided diagnosis and interventions. For example, for the planning of cardiac interventions it is necessary to identify standardized planes of the heart, *e.g.* short-axis and 2/4-chamber views [ALFV+18]. It also plays a crucial role for prenatal fetal screening, where it is used to estimate biometric measurements like fetal growth rate to identify pathological development [RPN12]. Moreover, the mid-sagittal plane, commonly used for brain image registration and assessing anomalies, is identified based on landmarks such as the AC and PC [AOL+19a]. Linking back to our case study from Chapter 3. An astronaut helping a fellow crew member undergoing an ultrasound screening will not necessarily have the medical expertise to find the appropriate

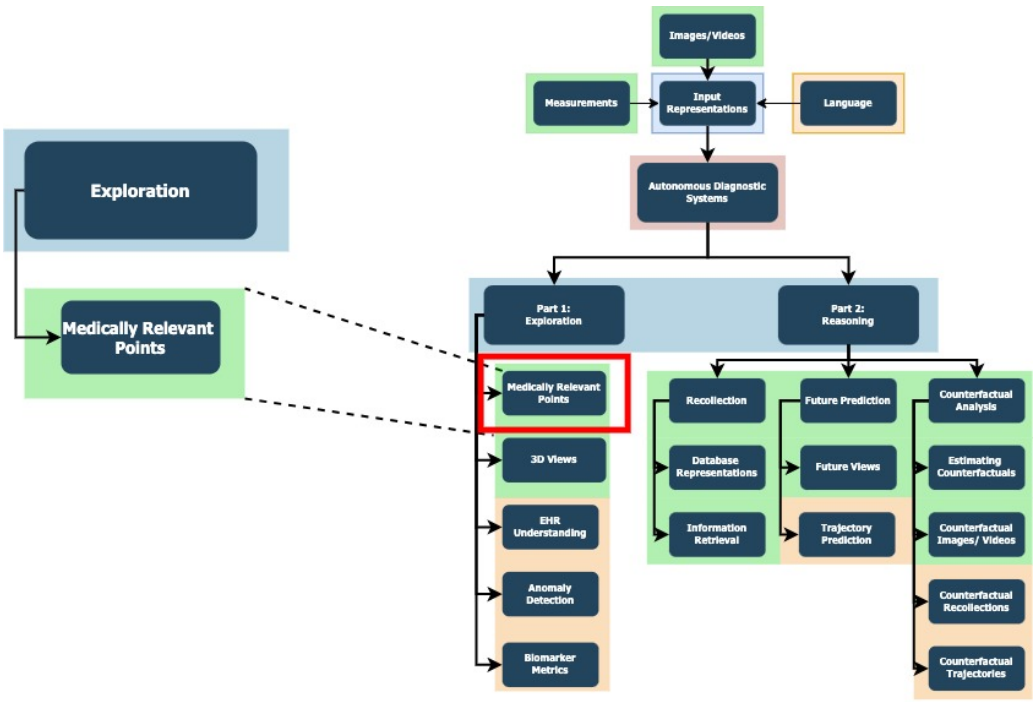


Figure 4.1: We first look into the exploration branch and how to identify points of medical interest

medical landmarks that would be informative for the ADS. As such, assuming a basic understanding of human anatomy the need arises for an automated method of guiding them to medically relevant points.

Manual annotation of landmarks is often a time-consuming and tedious task that requires significant expertise about the anatomy and suffers from inter- and intra-observer errors. Automatic methods on the other hand can be challenging to design because of the large variability in the appearance and shape of different organs, varying image qualities and artifacts. Thus, there is a need for methods that can learn how to locate landmarks with highest accuracy and robustness; one promising approach is based on the use RL algorithms [GGM+16, AOL+19a].

We present a novel Multi Agent Reinforcement Learning (MARL) approach for detecting multiple landmarks efficiently and simultaneously by sharing the agents' experience. We identify three main points in our approach: *(i)* We introduce a novel formulation for the problem of multiple landmark detection in a MARL framework; *(ii)* A novel collaborative Deep Q-Network (DQN) is proposed for training using implicit communication between the agents; *(iii)* Extensive evaluations on different datasets and comparisons with recently published methods are

provided (decision forests, Convolutional neural network (ConvNet), and single-agent RL).

4.1 Related Work

In the literature, automatic landmark detection approaches have adopted machine learning algorithms to learn combined appearance and image-based models, for example using regression forests [OBG+17] and statistical shape priors [GCLB15]. Zheng et al. [ZLG+15] proposed using two CNNs for landmark detection; the first network learns the search path by extracting candidate locations, and the second learns to recognize landmarks by classifying candidate image patches. Li et al. [LAC+18] presented a patch-based iterative CNN to detect individual or multiple landmarks simultaneously. Ghesu et al. [GGM+16] introduced a single deep RL agent to navigate in a 3D image towards a target landmark. The artificial agent learns to search and detect landmarks efficiently in an RL scenario. This search can be performed using fixed or multi-scale step strategies [GGZ+19]. Recently, Alansary et al. [AOL+19a] proposed the use of different DQN architectures for landmark detection with novel hierarchical action steps. The agent learns an optimal policy to navigate using sequential action steps in a 3D image (environment) from any starting point towards the target landmark. In [AOL+19a] the reported experiments have shown that such an approach can achieve state-of-the-art results for the detection of multiple landmarks from different datasets and imaging modalities. However, this approach was designed to learn a single agent for each landmark separately. In [AOL+19a] it has also been shown that performance of different strategies and architectures strongly depends on the anatomical location of the target landmark. Thus we hypothesize that sharing information while attempting simultaneous detection reduces the aforementioned dependency.

4.2 Proposed Method

We formulate the problem of multiple anatomical landmark detection as a multi-agent reinforcement learning scenario. Building upon the work of [GGM+16] and [AOL+19a] we extend

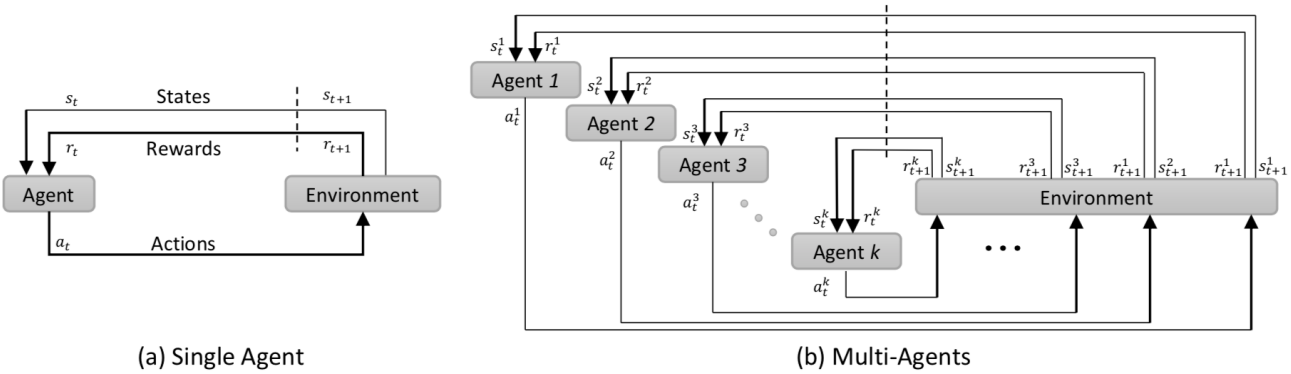


Figure 4.2: (a) A single agent and (b) multi agents interact within an RL environment.

the formulation of landmark detection as a Markov Decision Process (MDP), where artificial agents learn optimal policies towards their target landmarks, which defines a concurrent Partially Observable Markov Decision Process (POMDP) (co-POMDP)[GE15]. We consider our framework concurrent as the agents train together but each learns its own individual policy, mapping its private observations to a personal action [GEK17]. We hypothesize that this is necessary as the localization of different landmarks requires learning partly heterogeneous policies. This would not be possible with the application of a centralized learning system.

Our RL framework is defined by the *States* of the environment, the *Actions* of the agent, their *Reward Function* and the *Terminal State*. We consider the environment to be a 3D scan of the human anatomy and define a state as a Region of Interest (ROI) centered around the location of the agent. This makes our formulation a POMDP as the agents can only see a subset of the environment [JSJ95]. We define the frame history to be comprised of four ROIs. In this setup each agent can move along the x, y, z axis creating thus a set of six actions. The agents evaluate their chosen actions based on the maximization of the rewards received from the environment. The reward function is defined as the relative improvement in Euclidean distance between their location at time t and the target landmark location. In our multi-agent framework, each agent calculates its individual reward as their policies are disjoint.

During training, we consider the search to have converged when the agent reaches a region within 1mm of the target landmark. Episodic play is introduced in both training and testing. In training, the episode is defined as the time the agents need to find the landmarks or until they have completed a predefined maximum number of steps. In case one agent finds its landmark

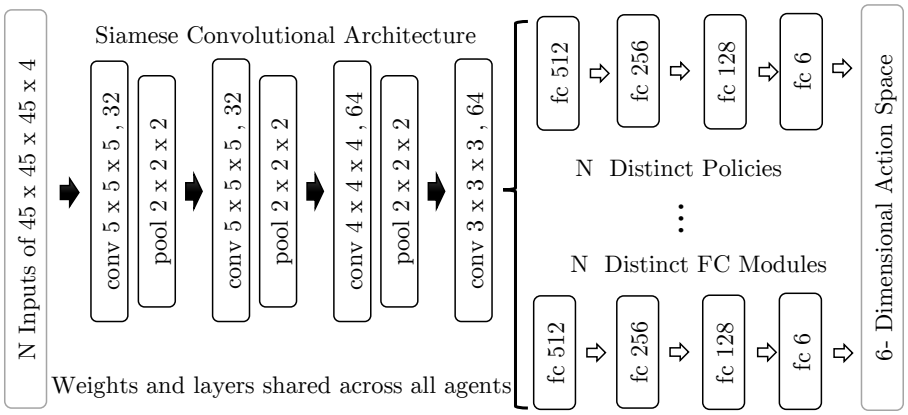


Figure 4.3: Proposed Colab-DQN for the case of two agents; The `convolutional` layers and corresponding weights are shared across all agents making them part of a Siamese architecture, while the policy making fully connected layers are separate for each agent.

before all others, we freeze the training and disable network updates derived from this agent while allowing the other agents to continue exploring the environment. During testing, we terminate the episode when the agent starts to oscillate around a position or exceeds a defined maximum number of frames seen in the episode similar to [AOL+19a].

4.2.1 Collaborative Agents

Previous approaches to the problem of landmark detection by [AOL+19a], [GGZ+19] and [GGM+16] considered a single agent looking for a single landmark. This means that further landmarks need to be trained with separate instances of the agent making a large scale application unfeasible. Our hypothesis is that the position of all anatomical landmarks is interdependent and non-random within the human anatomy, thus finding one landmark can help to deduce the location of other landmarks. This knowledge is not exploited when using isolated agents. Thus, in order to reduce the computational load in locating multiple landmarks and increase accuracy through anatomical interdependence, we propose a collaborative multi agent landmark detection framework called Colab-DQN. The following description will assume just two agents for simplicity of presentation. However, our approach scales up to K agents. For our experiments we show evaluations using two, three and five agents trained together.

A DQN is composed of three `convolutional` layers interleaved with `maxpool` layers followed by three `fully connected` layers. Inspired by Siamese architectures [BGL+93], in our Colab-DQN we build K DQN networks with the difference that weights are shared across the `convolutional` layers. The `fully connected` layers remain independent since these will make the ultimate action decisions constituting the policy for each agent. In this way, the information needed to navigate through the environment are encoded into the shared layers while landmark specific information remain in the fully connected ones. In Figure 4.3, we graphically represent the proposed architecture for two agents. Sharing the weights across the `convolutional` layers helps the network to learn more generalized features that can fit both inputs while adding an implicit regularization to the parameters avoiding overfitting. The shared weights enable indirect knowledge transfer in the parameter space between the agents, thus, we can consider this model as a special case of collaborative learning [GEK17] where collaboration and communication is implicit.

4.3 Experimentation

4.3.1 Dataset:

We evaluate our proposed framework and model on three tasks: (i) brain MRI landmark detection with 728 training and 104 testing volumes [JJ+08]; (ii) cardiac MRI landmark detection with 364 training and 91 testing volumes [dMDS+14] and (iii) landmark detection in fetal brain US with 51 training and 21 testing volumes. Each modality includes 7-14 anatomical ground truth landmark locations annotated by expert clinicians [AOL+19a].

4.3.2 Training:

During training an initial random location is chosen from the inner 80% of the volume, in order to avoid sampling outside a meaningful area. The initial ROI is $45 \times 45 \times 45$ pixels around

Method	AC	PC	RC	LC	CSP
Supervised CNN	-	-	-	-	5.47 ± 4.23
DQN	2.46 ± 1.44	2.05 ± 1.14	3.37 ± 1.54	3.25 ± 1.59	$\mathbf{3.66 \pm 2.11}$
Colab-DQN	$\mathbf{0.93 \pm 0.18}$	$\mathbf{1.05 \pm 0.25}$	$\mathbf{2.52 \pm 2.25}$	$\mathbf{2.41 \pm 1.52}$	3.78 ± 5.55

Table 4.1: Results in millimeters for the various architectures on landmarks across brain MRI and fetal brain US. Our proposed Collab DQN performs better in all cases except the CSP where we match the performance of the single agent.

the randomly chosen point. The agents follow an ϵ-greedy exploration strategy, where every few steps they choose a random action from a uniform distribution while during the remaining steps they act greedily. Episodic learning with the addition of freezing action updates for the agents that have reached their terminal state until the end of the episode is used, as detailed in Section 4.2.

4.3.3 Testing:

For each agent, we fixed 19 different starting points in order to have a fair comparison among the different approaches. These points were used for all testing volumes for each modality at 25%, 50% and 75% of the volume's size. For each volume the Euclidean distance between the end location and the target location was averaged for each agent for each of the 19 runs. The mean distance in mm was considered to be the performance of the agent in the specific volume.

Multiple tests have been performed using our proposed architecture. Comparisons are made against the performance on multi-scale RL landmark detection [GGZ+19], fully supervised deep ConvNet [LAC+18] as well as a single agent DQN landmark detection algorithm [AOL+19a]. In case of cardiac landmarks we compare with [OBG+17] that utilizes decision forests. Different DQN variations like the Double DQN or Duelling DQNs are not evaluated since their performance provides little to no improvement for the task of anatomical landmark detection as exhibited in [AOL+19a].

Even though our method can scale up to K agents given enough computational power we limited our comparison to the Anterior Commissure (AC) and the Posterior Commissure (PC) of

the brain; the Apex (AP) and Mitral Valve Centre (MV) of the heart; the Right Cerebellum (RC), Left Cerebellum (LC) and Cavum Septum Pellucidum (CSP) for the fetal brain. These are common, diagnostically valuable landmarks used in the clinical practice and by previous automatic landmark detection algorithms. For completeness and to facilitate future comparisons, we provide our performance comparison also for the training of three and five agents simultaneously. In Table 4.1, we show the performance of the brain MRI and fetal brain US landmarks using the different approaches. In Table 4.2 we exhibit the results for three and five agents trained simultaneously and the results for cardiac MRI landmarks.

4.4 Discussion:

As shown in Tables 4.1 and 4.2 our proposed method significantly outperforms the current state-of-the-art in landmark detection. p-values from a paired student-t test for all experiments were in the range 0.01 to 0.0001. We perform an ablation study by training instances of a single agent with double the iterations and double the batch size. The study has been conducted on the Cardiac MRI landmarks that have exhibited the biggest localization difficulties because of larger anatomical variations across subjects than observed in brain data. Our results confirm that the agents share basic information across them, which helps all of them perform their tasks more efficiently. These results support our hypothesis that the regularization effect from the gradients collected from the increased experience and knowledge of the multi-agent system

Landmark	**3 Agents**	**5 Agents**
AC	0.94 ± 0.17	0.98 ± 0.25
PC	0.96 ± 0.20	0.90 ± 0.18
Landmark 3	1.45 ± 0.51	1.39 ± 0.45
Landmark 4	N/A	1.42 ± 0.90
Landmark 5	N/A	1.72 ± 0.61

(a)

Method	**AP**	**MV**
Inter-Obs. Error	5.79 ± 3.28	5.30 ± 2.98
Decision Forest	6.74 ± 4.12	6.32 ± 3.95
DQN	4.47 ± 2.64	5.73 ± 4.16
DQN Batch $\times 2$	4.30 ± 12.07	5.01 ± 4.49
DQN Iterations $\times 2$	4.78 ± 13.87	5.70 ± 18.11
Colab-DQN	$\mathbf{3.96 \pm 5.07}$	$\mathbf{4.87 \pm 0.26}$

(b)

Table 4.2: (a)Multiple agent performance, training and testing were conducted in the Brain MRI; Landmarks 3,4,5 represent respectively the outer aspect, the inferior tip and the inner aspect of the CSP; (b) multi-agent performance on cardiac MRI dataset;

is advantageous. Furthermore, we created a single agent with doubled memory but due to the random initialization of experience memory, the agent failed to learn. In addition, as shown in Table 4.2(a), the inclusion of more agents leads to similar or improved results across all landmarks. It is interesting to note that even though we perform better in all landmarks, our approach can only match the performance of a single agent DQN for the CSP landmark. We theorize that this is due to the different anatomical nature of the RC, LC landmarks compared to the CSP landmark, thus the joint detection does not present an advantage. We chose to utilize the DQN in this chapter rather than existing policy gradient methods like Asynchronous Advantage Actor Critic (A3C) as the DQN is represented by a single deep ConvNet that interacts with a single environment. A3C use many instances of the agent that interact asynchronously and in parallel. Multiple A3C agents with multiple incarnations of such environments are computationally expensive. In future work, we will investigate the application of other methods for multiple-landmarks detection using either collaborative or competitive agents. Finally we ought to note that as the information passing is implicit and primarily effects the agents during training error accumulation and trajectory drift effects are minimized during inference. As such, if one agent experience a catastrophic divergence and fails to locate its target, other agents are not affected to an extent that would cause them to lose their targets as well.

4.4.1 Computational Performance:

Training multiple agents together does not only provide benefits in performance of landmark localization, it also reduces the time and memory requirements of training. Sharing the weights between the convolutional layers helps to reduce the trainable parameters by 5% in case of two agents and by 6% in case of three agents when compared with the parameters of two and three separate networks respectively. Furthermore, the addition of a single agent to our architecture reduces the required number of parameters by 6% compared to a single standalone agent. Due to the regularization effect that multiple agents have on their training and the implicit knowledge transfer, the training time our approach needs on average 25.000-50.000 less time

steps to converge compared with a single DQN and each training epoch needs approximately 30 minutes less than the training of 2 epochs in a separate single DQN (NVIDIA Titan-X, 12 GB). Inference is on par with a single agent at ∼20fps.

4.5 Chapter Summary

In this chapter we formulated the problem of multiple anatomical landmark detection as a multi-agent reinforcement learning scenario, we also introduced Colab-DQN, a Collaborative DQN for landmark detection in brain and cardiac MRI volumes and 3D US. We train K agents together looking for K landmarks. The agents share their convolutional layer weights. In this fashion we exploit the knowledge transferred by each agent to teach the other agents. We achieve significantly better performance than the next best method of [AOL+19a] decreasing the error by more than 1mm while taking less time to train and less memory than training K agents serially. We believe that a Bayesian exploration approach is a natural next step, which could be addressed in future work.

Since the time of writing and publication of this work, a series of methods have been proposed and built upon the algorithms of this chapter. [LRA20] directly extend the concept of MARL for anatomical landmark detection by building explicit communication channels between the agents. Meanwhile [YHH+21] incorporate time domain information by using an Recurrent Neural Network (RNN) backbone to their MARL algorithm. While the advances of anatomical landmark detection are numerous, certain main axes of research emerge. Methods solely concerned for with identifying the landmarks often opt of deep learning methods like the faster R-CNN [CLD+21]. On the other hand, robotics oriented applications that require both accurate detection of the landmarks and the trajectories of localization opt for reinforcement learning methodologies.

In this way we set our first building block towards an autonomous diagnostic system. In the next chapter we will be looking into how to contemplate and imagine the 3D structure of a 2D object we saw in our exploration.

Chapter 5

Imagining Objects

Having established methods to explore the environment and identify points of interest, we turn our focus to trying to image how the depicted 2D objects appear in 3D. Moreover we will try to segment away regions of interest in these 3D reconstructions. Understanding the 3D properties of objects is a vital part of intelligence and with this chapter we inch forward towards achieving this greater and difficult goal. This work is classified under the *Exploration* paradigm of the ADS as seen in Figure 5.1. This chapter is based upon [VBH$^+$20],published in the TIA workshop of MICCAI 2020 where the author conceived the method, planned and ran the relevant experiments.

X-Ray imaging is quick, cheap and useful for front-line care assessment and intra-operative real-time imaging (e.g., C-Arm Fluoroscopy). However, it suffers from projective information loss and lacks vital volumetric information on which many essential diagnostic biomarkers are based on. In this chapter we explore probabilistic methods to reconstruct 3D volumetric images from 2D imaging modalities and measure the models' performance and confidence. We show our models' performance on large connected structures and we test for limitations regarding fine structures and image domain sensitivity. We utilize fast end-to-end training of a 2D-3D convolutional networks, evaluate our method on 117 CT scans segmenting 3D structures from Digitally Reconstructed Radiographs (DRR) with a Dice score of 0.91 ± 0.0013.

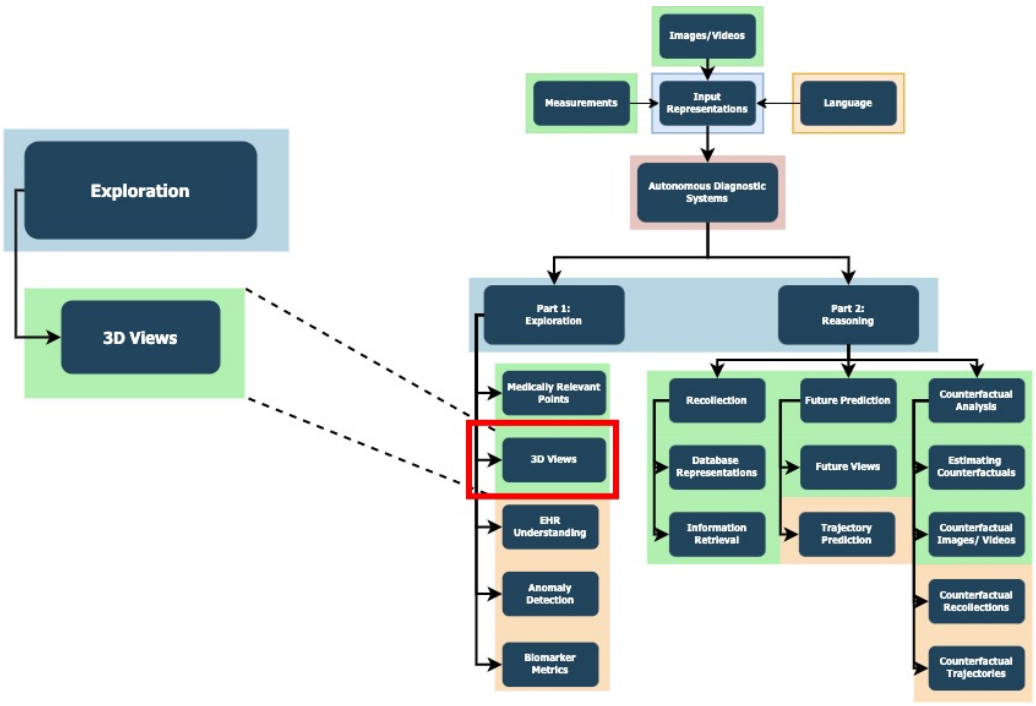


Figure 5.1: Subsequently we look into the exploration branch and how to imagine the 3D structure of a 2D object we observed

5.1 Introduction

Computed tomography (CT) scans provide detailed 3D information of patient anatomy that is vital to many clinical workflows. For many pathologies, accurate diagnosis relies heavily on information extracted from CT images and volumes [Rub14], e.g. biomarkers derived from 3D lung segmentations are used to characterize and predict Tuberculosis progression [VDVVdBDS03]. CT scans, however, are both time-consuming and expensive to perform, and are not always available at the patient's current location, resulting in delayed diagnosis and treatment. CT scans also present a higher risk to the patient due to increased radiation exposure over a typical Chest X-Ray (CXR). Meanwhile CXRs are routinely taken in the clinical practice at significantly decreased cost and radiation dosage while acquisition times are many orders of magnitude less than a CT scan.

Learning based methods have shown great potential for synthesizing structurally coherent information in applications where information is lost due to non-invertible image acquisition [HRRR18]. A primary example of such an application is CXR projection. As the human anatomy is locally well-constrained, a canonical representation can be adopted to learn the

anatomical features and extrapolate a corresponding 3D volume from a 2D projection view. This can be achieved by reflecting likely configurations, as they were observed in the training data, while inference is conducted by giving a sparse conditioning data sample, like a single projection.

We show how probabilistic segmentation techniques [BTC+19, BSK+19] can be extended with the ability to reconstruct 3D structure from projected 2D images. Our approach evaluates the potential of deep networks to invert projections, an unsolved problem of projective geometry and medical image analysis. We evaluate our method by reconstructing 3D lung segmentation masks and porcine rib-cages from 2D DRRs. We show that our approach works well for large, connected regions and test for limitations regarding fine, unconnected anatomical structures projected on varying anatomy and domain sensitivity across datasets. We further show how to adapt our methods to perform Unsupervised Domain Adaptation on NIH chest X-Rays. The proposed network is fast to train, converges within a few hours and predicts 3D shapes in real-time.

5.2 Related Work

Extracting 3D models from a single or multiple 2D views is a well-established topic in computer vision [LKL18, SWZ+18]. Earlier approaches included learning shape priors, and fitting the 3D shape model onto the 2D image. In [AVI+16, KW09], the authors attempt to reconstruct ribs by using, a priori known, statistical shape models. Both methods use a bi-planar approach as they utilize 2 orthogonal X-ray views. These methods do not generate a CT-like image, as they only deform a solid rib-like template.

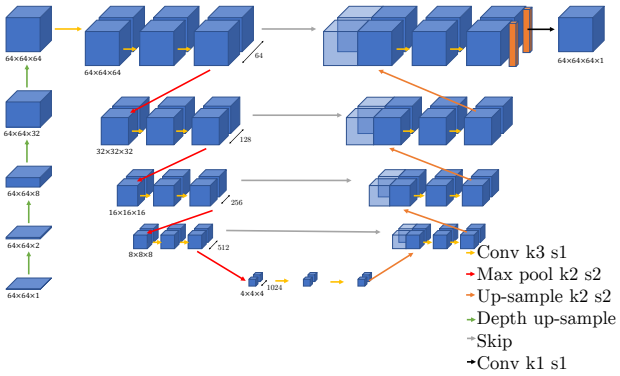
With the advances in deep learning, generative deep convolutional neural networks have been proposed to perform image generation in the context of medical imaging. In a recent work, parallel to ours, Ying et al. proposed X2CT-GAN [YGM+19] to synthesize full 3D CT volumes from 2D X-rays. Like [AVI+16, KW09], Ying et al. also use multiple views to create the 3D volume. However, instead of statistical shape models, Ying et al. uses Generative Adversarial

Network (GAN) to synthesize 3D CT volumes from 2D X-rays. As GANs are trained to approximate the probabilistic distribution of the training dataset implicitly, they are known to hallucinate "plausible" features. This is detrimental in cases of fine structures, e.g., bronchi, blood vessels and small lesions. In the case of vessel-like structures which are almost random in construction, a GAN will hallucinate a plausible structure that is highly probable from images in the training dataset, instead of generating a structure that's extrapolated from the input. Hence the resulting structures are of poor quality, often disconnected and non realistic.

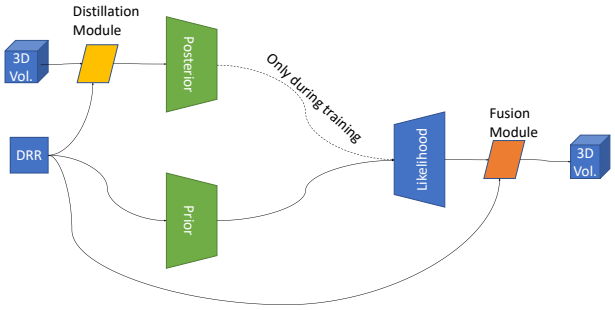
In [YYY+16] the authors reconstruct a 3D volume of an object from a 2D image of it. Contrary to X-Rays which can be thought of as the "shadow" of the object, [YYY+16] used as inputs 2D images of 3D structures, not their projections. Hence there was significantly less information loss than in the case of projections. [AFN17] attempts a similar task to ours but aims at decomposing the provided X-Ray image rather than reconstructing the CT volume. More aligned to our work, Henzler et al [HRRR18] creates 3D Volumes from 2D Cranial X-Rays. Their architecture is similar to ours, however, they only regress the 3D cranial structure, whereas we attempt to regress directly to CT Hounsfield units (HU). Furthermore we adopt a probabilistic technique while their model is fully deterministic.

5.3 Methodology

Adapting a known 2D or 3D architecture to be able to perform a task across dimensions is not a trivial task. Special consideration has to be given in the flow and augmentation of information. As projection is an information-destroying process our methods have to be able to deduce the lost information in order to revert the process. This can be achieved through appropriate pathways of information through the network. It is impossible to be entirely certain that the restored information is correct as projection is a many-to-one operation, thus we believe that a probabilistic approach can offer reasonable confidence intervals. We extend two base architectures to perform this task as they our outlined in Fig. 5.2.



(a) 2D to 3D U-Net, Blue Blocks indicate 3D Convolutions; Orange Blocks indicate Dropout Layers, *c.f.* Sect. 5.3.1



(b) PhiSeg[BTC+19] with proposed augmentations, *c.f.* Sect. 5.3.4

Figure 5.2: Two approaches for probabilistic 2D-3D un-projection.

5.3.1 2D to 3D MC-Dropout-U-Net

Our first proposed method extends the work of [BSK+19]. Inputs will be first transformed into three dimensional objects using the structural reconstruction module and then passed through a 3D U-Net [ÇAL+16]. The U-Net is equipped with dropout layers on the decoding path, which are kept active during inference to mimic stochastic behavior. Fig. 5.2a shows an overview over the proposed architecture.

5.3.2 Structural Reconstruction Module:

2D images can be considered as a 3D image with a "depth" of one. A series of five 3D transposed convolutional layers, with stride greater than 1 in the z-axis, is used to match the spatial dimensions of the volumetric 3D target. As opposed to bilinear up-sampling we propose to use transposed convolutions due to their theoretically better ability to learn more complex and non-linear image resizing functions [DV16]. The network at this stage contains a conceptual representation of the 3D properties of the object. As the 3D properties of the volume are yet to be fine-tuned by the subsequent 3D U-Net, the output of this layer does not hold human-understandable information of the 3D structure.

5.3.3 3D Segmentation:

With the input data in correct spatial dimensions, segmentation can be performed using a 3D U-Net [ÇAL+16]. Similarly to its well-known 2D counterpart, a 3D U-Net follows an encoding-decoding path with skip connections at each resolution. The network consists of four resolution layers; each consisting of two $3 \times 3 \times 3$ kernels with strides of $1 \times 1 \times 1$, followed by a $2 \times 2 \times 2$ max pooling with strides of $2 \times 2 \times 2$. Skip connections are used across the encoding and the decoding path, connecting same resolution levels, in order to propagate encoded features at each resolution. A dropout layer is added at the end of the decoder with a dropout probability of 0.6. These layers are kept active during inference as per the MC-Dropout methodology [GG16]. The network is then trained on 2D images with the respective 3D targets for segmentation and a binary cross-entropy loss.

5.3.4 2D to 3D PhiSeg

In [BTC+19] Baumgartner et al. introduce PhiSeg; a probabilistic segmentation algorithm able to capture uncertainty in medical image segmentations.

Phiseg is comprised of three modules; the prior, posterior and likelihood networks. The algo-

rithm is modeled after a Conditional Variational Auto-Encoder where the posterior and prior networks operate as the encoders producing a series of latent variables z in different resolution levels. The likelihood network operates as the decoder utilizing the different resolution latent variables sampled from a normal distribution to produce segmentations. It is worth noting that the posterior network takes as input the ground truth segmentation and hence it is only used during training. An auxiliary KL divergence loss between the distributions of the prior and the posterior is employed to steer the prior network to produce "good" latent variables.

We extend the previous method in three major ways aimed at controlling and augmenting the information contained in the DRR image.

1. Distillation Module:

We propose a "distillation" module that performs the inverse operation of the Structural Reconstruction Module and we add it as a pre-processing step of the posterior network. The ground truth image is passed through a series of convolutional layers to "distill" its 3D information to a 2D representation. The goal of this "distillation" is to build an informative lower dimensional representation of our 3D volume. The resulting feature maps are concatenated with the input DRR image and passed through the posterior network. Contrary to the aforementioned 2D-3D U-Net PhiSeg is modeled after a VAE, hence the encoded latent distribution is highly susceptible to noisy inputs. In order to avoid the encoding of noise that would change the characteristics of our distribution we chose to work on 2 dimensions during the encoding phase rather than in 3. We would like to note that a fully 3D PhiSeg with a Structural Reconstruction Module as in the 2D-3D U-Net was evaluated but its training was unstable.

2. 3D Likelihood network:

We extend the likelihood network to perform 3D reconstruction. The latent variables that the prior/posterior networks produce are transformed into 3D vectors and used as inputs for the likelihood network. We extend the latent vectors using vector operations rather than learning

an augmentation to decrease the computational load of the the network. The series of latent variables are then passed through 3D decoder network, sharing the same architecture as the decoder path of the deterministic 3D U-Net.

3. Fusion Module:

Our next extension of PhiSeg comes in the form of a fusion module similar to [HRRR18] at the end of the likelihood module. Contrary to [HRRR18] our fusion method is fully learned by the model. Features extracted from the input DRR image x are concatenated to the output s of the likelihood network and convolved together to produce s' which serves as the final output of the network. The intuition behind this module lies with the assumption that PhiSeg will be able to reconstruct the overall structure but may lack details, thus the input DRR image is passed through a convolutional layer to extract relevant features which are then used in conjunction with the proposed segmentation s. We also note that the fusion module is not included in the 2D-3D U-Net as the direct skip connections of the model satisfy the flow of information that the fusion module aims at creating.

4. Unsupervised Domain Adaptation:

Finally we propose a new augmentation of PhiSeg aimed at performing Unsupervised Domain Adaptation through self supervision. We chose the task of reconstruction as an auxiliary task in accordance with [STDE19] since it is semantically close to our target segmentation task. To this end we make a new copy of the prior/posterior and likelihood networks that share the weights of the aforementioned modules. We train the resulting model for both segmentation and reconstruction in parallel. Hence the shared encoding paths learn to extract useful information from both domains. In section 5.4.3 we exhibit results using this technique to segment lungs from NIH X-Rays.

5.4 Experiments and Results

For our experimentation we focus on two tasks, segmentation and volumetry. Two datasets have been used: 60 abdominal CT images of healthy human patients (*Exp1*), and 57 CT porcine livers [KVSe14] (*Exp2*). Both datasets are resampled to isotropic spacing of 1mm$\times$1mm$\times$1mm. DRRs $\mathbf{p}$ were then generated by projecting the 3D volume on the DRR plane $\mathbf{p} = \mathbf{Mf}$ according to:

$$p_{i,j} = \sum_{i,j,k} m(i,j,k) f(i,j,k)$$

where $\mathbf{f}$ is the voxel density vector and $\mathbf{M}$ is the projection matrix calculated using the Siddon-Jacob's Raytracing algorithm [Jia10]. The synthetic X-ray images of the thorax and porcine abdomen are taken at a fixed distance of 2m and 1m respectively from the CT volume's isocenter, pixel spacing is $0.51mm$. Images contain 512×512 pixels, which in this particular configuration, aligns the DRR image and CT volume spatially in pixel space. Both images were then downsampled to 64×64 for network training, with the CT volume target centre cropped to preserve spatial alignment with respect to the DRR input. A third dataset (*Exp3*), obtained from the NIH Clinical Center, was used for a qualitative ablation study. 100 random chest X-ray images from the ChestX-ray8 [WPL$^+$17] dataset were selected. No ground truth is present for this experiment.

5.4.1 *Experiment 1*: Compact Structures

The first experiment assesses the network's ability to segment large connected regions. The thoracic CT dataset was used, with data split; 50 volumes for training and 10 for testing. Annotations were manually made to create ground truth masks for the lung structures. As the lungs appear much darker than other body structures, direct regression to the CT volume is a comparable ground truth target to the manual segmentation masks.

All networks are trained using the Adam-optimizer with an initial learning rate of 1×10^{-4} and a batch size of four. The resulting segmentations were post processed by thresholding based on

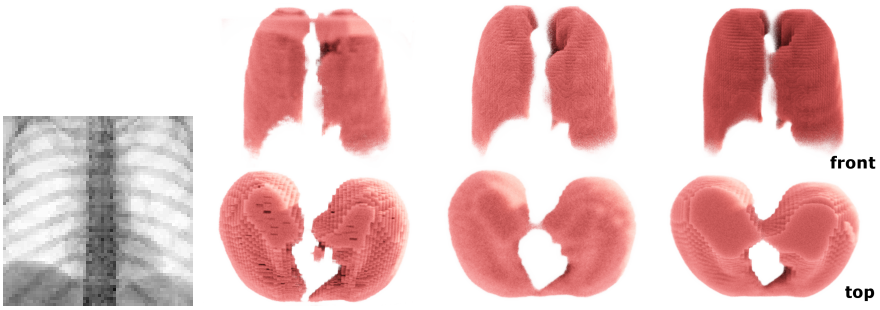
Method	Exp.1 Volume Ratio	Exp.1 Dice	Exp.2 Dice
Det. U-Net	0.96	0.86	0.41
2D-3D PhiSeg	0.92 ± 0.12	$\mathbf{0.91 \pm 0.01}$	0.46 ± 0.05
2D-3D PhiSeg w/o fusion	1.31 ± 0.22	0.81 ± 0.05	0.45 ± 0.07
2D-3D U-Net Dropout	0.91 ± 0.01	0.90 ± 0.01	$\mathbf{0.48 \pm 0.03}$
2D-3D U-Net Dropblock	$\mathbf{0.97 \pm 0.012}$	0.83 ± 0.007	0.36 ± 0.12

Table 5.1: Average Dice score and Volume Ratio ($\frac{Predicted}{True}$) of lung and porcine segmentations compared to manually generated 3D ground truth. Exp.2 shows the performance of our methods when the target task is to reconstruct fine 3D structures. In the first experiment we compare the predicted volume in 3D and compare it to the ground truth one. With a perfect score of 1, the volumes are calculated by summing the voxels of the volumes in question. Moreover the dice scores (higher being better) aim to assess the correct localization and fine structures of our predictions.

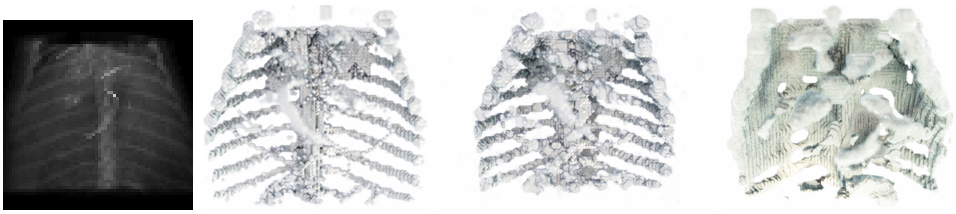
their pixel intensity values followed by median filtering with a kernel size of 3×3 to eliminate sparse noise.

Table 5.1 shows the average Dice similarity coefficient (DSC) for the predicted volume compared to the target volume. Dice accuracy for both approaches give equivalent performance. Table 5.1 also exhibits the ratio between the predicted volume of the lungs and the ground truth. This is is achieved by counting the pixels that lay inside the segmented volume. In terms of quantitative evaluation our deterministic model achieves high Dice score. Meanwhile our dropout and dropblock probabilistic approaches provide us with an on par or better performance to the deterministic method. The variance exhibited on a per sample basis is 0.02 on dice score and 0.03 on the ratio of lung capacity. The probabilistic method provides us with more informative lower and upper bound. As the process of projection inherently destroys information, it is our belief that providing informed upper and lower bounds for our metrics is a more suitable approach.

Furthermore a version of Phi-Seg without our proposed fusion module was evaluated and noticed a significant increase in the variance of our measurements as well as degraded performance. This observation is in accordance with our hypothesis that the fusion model inserts high level details to our proposed segmentation. Qualitative examples are shown in Figure 5.3(a).



(a) Lung Segmentation with 2D-3D Unet; Left-Right: Input DRR; 3D Ground Truth; 3D Prediction 2D-3D-Unet; Prediction 2D-3D PhiSeg



(b) Porcine Rib Cage Segmentation with 2D-3D Unet; Left-Right: Input DRR; 3D Ground Truth; 3D Prediction 2D-3D-Unet; Prediction 2D-3D PhiSeg

Figure 5.3: Reconstructed Samples for Experiments 1,2. We use [KPB12] to enhance depth perception in the 3D figures. Interestingly we observe that both methods are quite accurate in predicting the lung volumes, while the PhiSeg struggles with the resolution of fine structures in the porcine rib cage.

5.4.2 *Experiment 2*: Fine Structures

In order to test for limitations and to evaluate the network's ability to segment fine structures we aim to segment the ribcage with the publicly available porcine CT dataset [KVSe14]. Porcine ribs are smaller and finer than human ribs and they project largely on different anatomy (stomach and liver). This data has higher resolution and anatomical focus than the dataset in *Exp 1*, which serves as additional robustness test. The dataset consists of 58 volumes and has been split into 48 volumes for training and 10 for testing. Automated thresholding via pixel intensity was used to provide a manual ground truth from the 3D volumes. Known Hounsfield units (HU) for bones in CT have been used to define this threshold (+1800 to +1900 HU). The network has been trained with a binary cross-entropy loss, using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of four. Similarly to *Exp1*, the input to the network is a two dimensional DRR image while the segmentation target is the 3D segmentation mask.

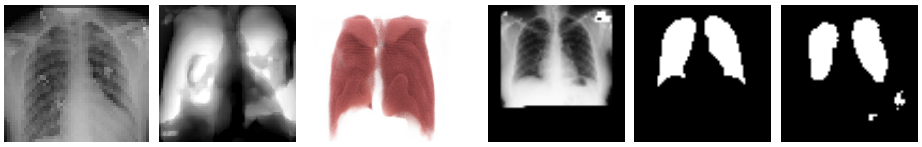
The resulting segmentations achieved an average Dice score of 0.41 in the deterministic case.

Meanwhile our probabilistic approaches were on par or better than the deterministic, achieving a Dice of 0.48 ± 0.03, while providing us with a more informed inference. We note that the difference between our PhiSeg model with and without the fusion module is smaller but still present. We believe this is due to the much harder task of segmenting fine structures across dimensions. Furthermore the Dice score is highly influenced by small outliers caused by noise and a blurry reconstruction of the spine as well as a slight misregistration between the predicted and ground truth volumes. Qualitative results are provided in Fig. 5.3(b). Note that small and fine structures as the tips of the ribs are reasonably well formed and shown in the predicted volume.

5.4.3 *Experiment 3*: Domain Adaptability

In order to evaluate the nature of knowledge acquired by the network, and to test potentially limited domain invariance, the network that has been used and trained for *Exp1* was evaluated with chest X-ray images from the NIH chest X-Ray dataset [WPL+17]. In addition we evaluate our UDA method on the Montgomery Chest X-Ray dataset that is comprised of 2D chest X-Rays with corresponding 2D segmentations.

As can be seen from Figure 5.4(a), where the lungs are semi-occluded by an imaging artifact, our network produced the underlying 3D segmentation. This observation signifies that the network learns to reconstruct the anatomy rather than learning a mean lung segmentation. Without corresponding CT volumes, it is not possible to quantitatively evaluate the performance of the network. However, qualitative assessment of 91 subjects shows robust performance of our approach. In the Montgomery dataset the resulting 3D segmentation perimeters are unknown. Thus, we learn a projection to 2D and then compare to the ground truth, resulting in a dice score of **0.77** when we optimize towards the main DRR-CT task and **0.86** when we optimize towards the UDA. It is important to note that information is lost during the projection from 3D to 2D during the evaluation period, which explains the decreased performance. In Figure 5.4(b) we show a selected example from the UDA algorithm.



(a) UDA on NIH Chest X-Ray Dataset; Left-Right: Input X-Ray; Mean of predicted volume across z-axis; 3D reconstruction of volume.

(b) UDA on the Montgomery X-Ray Thorax Dataset; Left-Right: Input X-Ray; 2D Ground Truth; 2D-3D PhiSeg volume projected onto 2D

Figure 5.4: Examples from Experiment 3.

5.5 Discussion

As shown in *Experiment 1*, our proposed method achieves good Dice scores for 3D lung segmentation while providing informative uncertainty as lower and upper bounds of the volume and dice score. To the best of our knowledge, this is the first probabilistic method to perform cross-modality 3D segmentation by unprojecting 2D X-ray images with acceptable performance. *Experiment 2* and *Experiment 3* have been designed to test expected limitations. In *Experiment 2* we observe that the prediction of fine structures can work, but with varying performance for either of the methods. *Experiment 3* shows that our method has promising domain adaptation properties. However, fine-tuning and calibration will be needed for applications.

5.6 Summary

In this chapter we have introduced simple methods to perform probabilistic 3D segmentation from a projective 2D X-ray image. Our networks are data efficient as they have been trained with approximately 60 training DRR-CT pairs and time efficient as they converge within ~ 2 hours. In future work we could explore the capabilities of our approach for the reconstruction of vessel trees, e.g. coronary arteries from C-Arm Fluoroscopy. We expect that such reconstructions can be well suited to accurately initialize the registration of pre-operative scans.

Having established methods for finding points of interest in medical images and understanding the 3D properties of 2D depictions of objects we can move forwards to recalling our past experiences that could help us reason about the information we extracted from our observations.

Part II

Reasoning

Chapter 6

Reasoning about the Future

We now turn our focus to applications relating to the central notions of this thesis. To make our navigation in medical environments and the reasoning about the potential pathologies more effective we need the ability to assess future developments, that is the reason we have dedicated an entire sub-branch of our ADS on future outcome prediction as seen in Figure 6.1. Even though our predictions cannot be pinpoint accurate, a probabilistic view of possible futures that are the causal consequence of the gathered facts and information can greatly help to establish possible courses of action. In relation to our case study Chapter 3, we envision our deployed ADS system to run simulations of possible outcomes and assessing the paths it needs to take in order to aid the astronaut in need. This chapter is based upon [VRRK20],currently under review in JMLR, where the author conceived the main ideas, developed and ran the experiments for the proposed methods and tools. The paper form of this chapter is currently under review at the Journal of Machine Learning Research.

6.1 Introduction

In many everyday scenarios we make causal predictions to assess how situations might evolve based on our observations and experiences. Machine learning has not been developed to this level yet. However, automated, causally plausible predictions are highly desired for critical

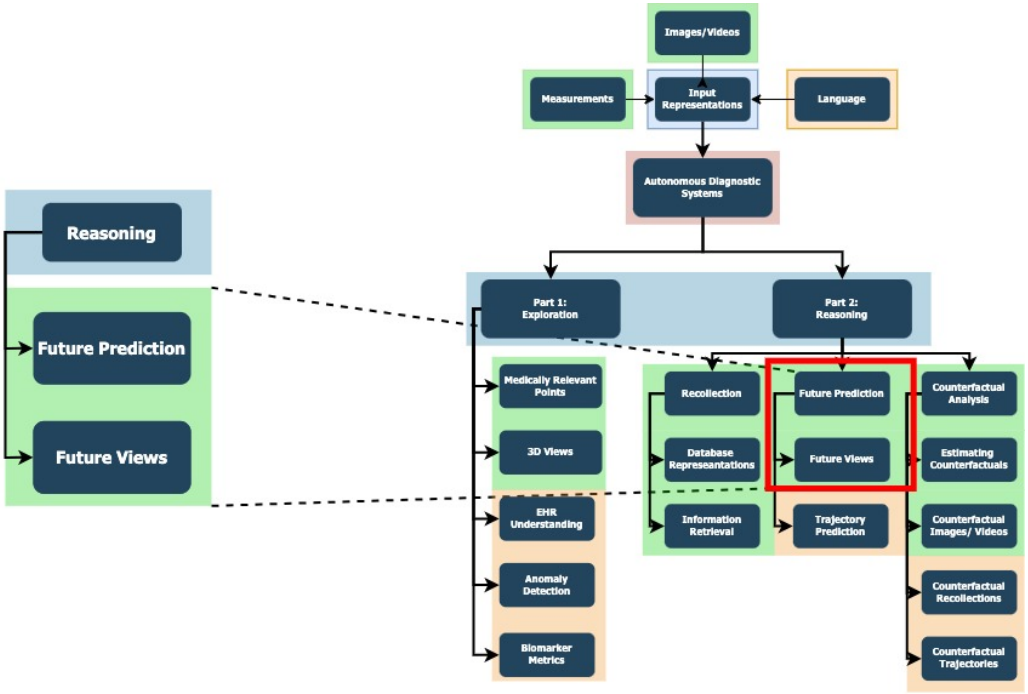


Figure 6.1: In this chapter we will be looking into the ability of our ADS to reason about how a scene might evolve visually

applications like medical treatment planning, autonomous vehicles and security. Recent works have contributed machine learning algorithms for the prediction of the future in sequences and for causal inference, *e.g.*, [KTY+18]. One major assumption that many approaches implicitly adopt, is that the space of the model representation is a flat Euclidean space of N dimensions. However, as shown by [AHH18], the Euclidean assumption leads to false conclusions as a model's latent space can be better characterized as a high dimensional curved Riemannian manifold rather than an Euclidean space. Furthermore, the Alexandrov-Zeeman theorem [Zee64, KK14] suggests that causality requires a Lorentzian group space and advocates the unsuitability of Euclidean spaces for causal analysis.

In this chapter, we present a novel framework that changes the way we treat hard computer vision problems like the continuation of frame sequences. We embed information on a spatio-temporal, high dimensional pseudo-Riemannian manifold - the MST - and utilize the special relativity concept of light cones to perform causal inference. We focus on temporal sequences and image synthesis to exhibit the full capabilities of our framework.

In summary our contributions are:

- We extend representation learning to spatio-temporal Riemannian manifolds that follow the ideas of the MST while being agnostic towards the used embedding architecture and the prescribed task.

- We introduce a novel utilization of the concept of light cones and use them for convincing frame synthesis and plausible prediction of future frames in video sequences.

- We provide theoretical guarantees about the causal properties of our model and demonstrate a causal inference framework.

6.2 Related Works

High dimensional Riemannian manifolds for machine learning are utilized by a few major works. [AHH18] show evidence that more general Riemannian manifolds characterize learned latent spaces better than an Euclidean space. Their work however, utilizes generators that have been trained under an Euclidean assumption. Contrary to that, [NK17] introduce the use of a Poincaré ball for hierarchical representation learning on word embeddings, showing superior performance in representation capacity and generalization ability while employing a Riemannian optimization process. [NK18] extend the previous work to a Lorentzian manifold as this offers improvements in efficiency and stability of the distance function. In this chapter we accept the argument made by Nickel *et al.* but extend it as we argue in Section 6.3.1 that causal inference requires a Lorentzian group space as pointed out by [Zee64].

[GBH18] embed word information on a Poincaré ball and form entailment cones. The authors propose to work with Directed Acyclical Graphs (DAG) and strive for non overlapping cones in a Poincaré ball. In contrast to this, we encourage overlapping light cones in a Lorentzian manifold to model future events.

[SWKMM15] use a space-time idea similar to ours but interpret the time axis as a ranking rather than as temporal information. Their method is intended for dimensionality reduction and does not generate further samples, or considers causal relationships between sampling points.

Finally, [MLM+19] train a Variational Autoencoder (VAE) constrained to a Poincaré ball while also employing the appropriate Riemannian equivalent to a normal distribution as well as Riemannian optimization. We consider this work as the closest related since it is the only approach that has shown good performance in the image domain.

In the Computer Vision focused field of future frame prediction for video sequences, [KTY+18] propose the causal InfoGAN which, however, lacks theoretical guarantees for causal abilities. [JEEL19] aims at predicting the probabilistic bottlenecks where the possible futures are constrained instead of generating a single future. Similarly, we are not attempting to predict a single future, rather we predict all plausible futures in a way that naturally enables us to identify all probabilistic bottlenecks; see Section 6.3.4.

In other works concerned with video continuation, *e.g.*, [MCL16, VPT16a] use CNNs to regress future frames directly, while [VYH+17] introduce an LSTM utilizing the difference Δ between frames to predict motion. Further works include the use of optical flow [LLLG18] or human pose priors [VYZ+17]. The autoregressive nature of these methods results in accumulated prediction errors that are propagated through the frames the further a sequence is extended. Beyond a few frames, these approaches quickly lose frame-to-frame coherence.

In order to mitigate these limitations, works like [VPT16b] propose generative models to predict future frames and [TLYK18a] offers a generative model that disentangles motion and content. Neither can infer the causal implications of their starting positions. [SLV21] perform video action prediction using a hyperbolic latent space, however, their method does not tackle any causality related questions nor is it free of autoregressive error compounding.

Finally, [LTA+20] is the closest work to ours. [LTA+20]'s aim is causal discovery, which they perform by reducing the problem of causal link prediction to pictorial structure prediction; *i.e.*, their method predicts deformable links between key points of objects limiting the application range.

6.3 Theoretical Formulation

6.3.1 On the choice of space

In his seminal 1964 work, E.C. Zeeman [Zee64] makes the case that the causality group $\mathcal{R}M$ that arises from the concept of partial ordering in a MST implies an inhomogeneous Lorentz group as the symmetry group of $\mathcal{R}M$. We highlight the explicit mention of Zeeman on the unsuitability of an Euclidean topology to describe $\mathcal{R}M$ due to its local homogeneity, which does not arise in $\mathcal{R}M$. In [KK14] the authors prove that from observable causality we can reconstruct the MST. Hence, we are in a position to argue that the use of a MST for embeddings, which belongs to the inhomogeneous Lorentz group, would reinforce causal inference capabilities.

We extend [NK18] and argue that the use of the Lorentzian manifold, which coincides with the MST, is both more efficient as an embedding as well as enabling causal arguments,

6.3.2 Minkowski Space-Time and Causality

Mathematically a space can be described by its metric, which defines the way the inner product of two vectors in this space is determined, *i.e.*, the way we calculate distances. Consequently, the inner product $\langle ., . \rangle_\eta$ of two vectors a and b in $1+3D$ Minkowski space-time can be defined as

$$\begin{aligned} \langle a, b \rangle_\eta &= \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} a_\mu \eta_{\mu\nu} b_\nu \\ &= -a_0 b_0 + a_1 b_1 + a_2 b_2 + a_3 b_3, \end{aligned} \tag{6.1}$$

where the coordinate 0 is understood to be the time coordinate.

One of the consequences of endowing the latent space with a Minkowski-like metric is the emergence of causality in the system. This property can be more readily seen by employing the

concept of *proper time*. Given a manifold $\mathcal{M}$ endowed with a Minkowski metric $\eta_{\mu\nu}$, we define the proper time τ. This is the time measured by an observer following along a continuous and differentiable path $\mathcal{C}(s)$ parametrized by $s \in [0,1]$ between two events $\{x, y\} \in \mathcal{M}$ such that $\mathcal{C}(0) = x$, $\mathcal{C}(1) = y$,

$$\tau_{\mathcal{C}} = \int_{\mathcal{C}} \sqrt{-\sum_{\mu,\nu} dx_\mu dx_\nu}. \tag{6.2}$$

In order to ensure $\tau \in \mathbb{R}$, we require $\sum_i dx_i^2 \leq dx_0^2$, where $i \in 1, 2, ..., d$. Therefore, the rate of change $|\mathbf{dx}|/d\tau$ in the spatial coordinates is capped by the time evolution of the system. In other words, there exists a maximum speed limit which $\mathcal{C}$ must obey at every point. Further, it means that there exist pairs of space-time points x, y which cannot be possibly connected by a valid path $\mathcal{C}$, lest $\tau \notin \mathbb{R}$. In order to describe this phenomenon we borrow the concept of a *light cone* from special relativity. The set of solution paths $\{\mathcal{C}_0(s)\}$ such that $\mathcal{C}_0(0) = (t_0, \mathbf{x_0})$ and $\tau_{\mathcal{C}_0} = 0$ describe the fastest any particle or piece of information can travel within the system starting from $(t_0, \mathbf{x_0})$. This boundary is known as the light cone, and is such that $\partial\mathcal{R} = \{\mathcal{C}_0(s)\}$, where $\mathcal{R}$ is the causal region of the point $(t_0, \mathbf{x_0})$. Every space-time point $x \in \mathcal{R}$ is said to be within the light cone. As shown by Eq. (6.2), no valid path $\mathcal{C}(s)$ can cross $\partial\mathcal{R}$. Thus, two space-time points can only influence each other if they lie within each other's light cone, that is, if they can be connected by a valid path $\mathcal{C}$. The region $\mathcal{R}$ splits into two disjoint sets: $\mathcal{R}^+$ and $\mathcal{R}^-$. $\mathcal{R}^+$ lies within the future light cone of a particle at time t_0, and thus includes all of the points $(t_1, \mathbf{x_1}) \in \mathcal{R}$ such that $t_1 > t_0$. Conversely, $\mathcal{R}^-$ includes the points $(t_2, \mathbf{x_2}) \in \mathcal{R}$ such that $t_2 < t_0$ and characterizes the past light cone of a particle at time t_0.

If we have two space-time vectors $x = (t_0, \mathbf{x_0})$ and $y = (t_1, \mathbf{x_1})$ we can describe their relation as *timelike* when $\langle x, y\rangle < 0$, *spacelike* when $\langle x, y\rangle > 0$ and *lightlike* when $\langle x, y\rangle = 0$. A timelike position vector lies within the light cone of a particle at the origin of the system. A spacelike vector lies outside of it, and a light-like vector lies exactly at its edge. One can then generalize this idea beyond the origin, and thus compute the inner product of the difference between two space-time vectors $x - y \equiv (\Delta t, \Delta\mathbf{r})$, *i.e.*,

$$\langle y - x, y - x\rangle = -\Delta t^2 + |\Delta\mathbf{r}|^2. \tag{6.3}$$

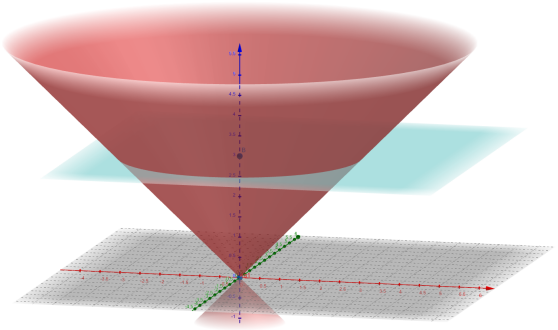
Hence, when the separation of the vectors x and y is time-like, they lie within each other's causal region. In that case we can argue that there is a path for event x, that belongs in the model that defines the latent world of represented data, to evolve into event y within a time period Δt. Thus, by constructing the light cone of an initial point x we can constrain the space where the causally resulting points may lie. We can then see that this mathematical construction of the latent space naturally enforces that the velocity of information propagation in the system be finite, and that an event can only be influenced by events within its past light cone, *i.e.*, the model is causal. By mapping this into a machine learning perspective we argue that in a latent space that is built to follow the MST metric an encoded point can then be used to create a light cone that constrains where all the causally entailed points may be encoded to or sampled from.

6.3.3 On Intersecting Cones - Association

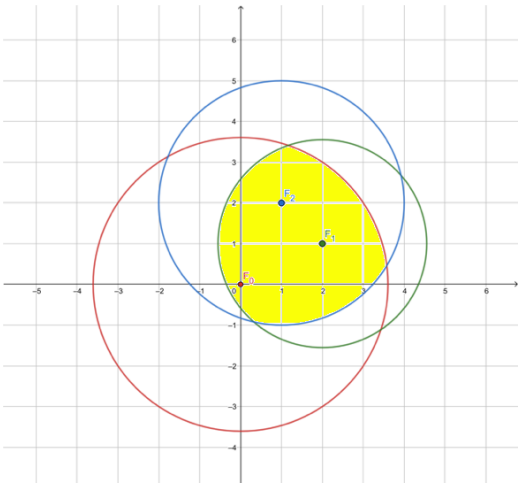
A light cone can be constructed with each point of the latent space as its origin. Consider point x_0 to be an initial point derived from, for example, an encoded frame f_0 from a video sequence: by constructing the light cone C_0 around x_0 we are able to deduce where the various causally related x_{0+t} points might lie. By setting t to be a specific time instance, we are able to further constrain the sub-space to points that lie inside of the conic section. These points are causally plausible results from point x_0 within the time t. Geometrically, we can visualize this as a plane cutting a cone at a set time. We visualize this in Figure 6.2a.

A second point x_1 that lies inside the light cone of x_0 can be derived from an encoded frame f_1. Similar to x_0 we construct the light cone C_1 whose origin is x_1. We then define the conic intersection $CS = C_0 \cap C_1$. Following the causality argument, we deduce that the enclosed points in CS are causally related to both x_0 and x_1 as they lie in the light cones of both. In addition, by constraining the intersecting time plane, we constrain the horizon of future prediction.

Consequently, we propose Algorithm 1 as a method of future frame prediction using light cones on a Minkowski space-time latent space. We graphically illustrate Algorithm 1 in Figure 6.2b.



(a) Visualization of the emerging structure of a light cone. The intersecting plane at point $z = 3$ signifies the 2-dimensional feature space at time 3. The interior of the cone subspace contains all possible frames given a original video frame at point $z = 0$.



(b) Visualization of the intersecting cones algorithm. The subspace marked in yellow contains the points that are causally related to points $F_{0,1,2}$.

Figure 6.2: Visual aids for the proposed algorithm. Note that for visualization purposes we are exhibiting a $1+2$ dimensional Euclidean space rather than a high dimensional Riemannian manifold.

Algorithm 1 Future Prediction using Intersecting Light Cones

Input: Frame Sequence F ; Queried Time T
Output: Predicted Frame

for $t < T$ **do**
  $Mf_t \leftarrow MinkowskiEmbedding(f_t)$
  $C_{Mf_t} \leftarrow LightCone(Mf_t)$
  if $t > len(F)$ **then**
    $Samples_{Mf_t} \leftarrow sample(C_{Mf_t})$
    $Mf_{t+k} \leftarrow choose(Samples_{Mf_t})$
  end if
  $CS \leftarrow intersection(C_{MF})$
  $f_{out} \leftarrow choose(sample(CS))$
  Predicted Frame $\leftarrow Decoder(f_{out})$
end for

6.3.4 On the Entropy and the Aperture of Cones

When considering the intersection of the cones in Algorithm 1 it is vital to examine the aperture of the cone at time T. For simplicity, we assume that the gradient of the side of the cone is 45° for all cones. However, such an assumption implies that each frame and hence each cone evolves with the same speed and can reach the same number of states at a given time. For real world scenarios this is not necessarily true as, for example, the possible states in $t+1$ for a ball rolling constrained by rails are less than a ball rolling on a randomly moving surface. Hence, the actual gradient of the cone depends on the number of states that are reachable from the state depicted

in frame t. This quantity is also known as the thermodynamic entropy of the system. It is defined as the sum of the states the system can evolve into. Calculating the thermodynamic entropy of a macro-world system as in a real world dataset is not trivial and we are not aware of any appropriate method to compute this at the time of writing. Hence, we are forced to make the aforementioned assumption of 45° of cone gradient. We note that the concept of information entropy commonly used in computer science is distinct to thermodynamical entropy and a map between the two is non trivial.

However, given a frame sequence F, a set of counter example frames CF and following Algorithm 1 but omitting the sampling steps, it is possible to build more accurate light cones in a contrastive manner. Hence, it is possible to acquire a proxy for the thermodynamic entropy of the system. We note that the proxy can only be accurate to a certain degree as any frame sequence is not able to contain enough information to characterize the full state of the world. We leave further study of this phenomenon to future work.

6.3.5 Intervention on Minkowski Space Time

Given an MST where we have embedded our beliefs about the world we are able to further perform hypothetical causal interventions. These interventions come into the form of choosing specific paths as in step *choose* of Algorithm 1. Given a heuristic $\mathcal{H}$ which is application dependent and could even be parametrized by a neural network, we enforce a path that represents the $do(Y)$ operation. Performing causal interventions by setting paths through the MST gives us the additional advantage of having complete control of the process, allowing the interventions to be fully interpretable. The difference compared to regular associations is that we enforce a specific event to happen by choosing a point in our latent space and then construct the appropriate conical structures. In addition, we note that this process affects the probabilities of future causally linked events only locally rather than globally as associations do, this is in accordance to the general intuition of causal intervention, as nicely explained in [GMD+20]. In Figure 6.3a we illustrate this method.

6.3.6 Counterfactuals on Minkowski Space Time

As explained in Section 2.9, counterfactual inference can be achieved either following the three step process laid out by [Pea09] or via the difference of the average treatment effect as discussed by [Rub05]. In this section we will show how both these approaches could be interpreted in a graphical way inside the MST. In Figure 6.3b, we exhibit a visualization of a sample counterfactual analysis of $\hat{X}$ with X being the factual prior. In other words, given the factual knowledge X we set up the future and past lightcone of point X. This is equivalent to the *Abduction* step by Pearl. Our method and probabilistic abduction are equivalent as in both cases we constrain the possible subsequent events and update their probabilities given the conditioning one. Next, we identify the other conditioning variables B and based on that we perform the $do(Y')$ operations as required, fulfilling the *Action* step. At this point we are able to analyze the resulting modified lightcones and identify the probability of the counterfactual query $\hat{X}$.

Moreover, we are able to perform a Rubin inspired average treatment analysis. Such an analysis is achieved by identifying the points inside the conditioning factual event's lightcone representing the intended treatment and counterfactual treatment and then analyzing the required counterfactual event queries.

Most counterfactual queries assume a partial ordering in time and ask questions about alternative timelines where different events have occurred in the past. Contrary to other methods where time is a latent variable dealt with implicitly, our method explicitly includes time in the process of causal inference permitting for more accurate conclusions to be drawn.

6.3.7 Step-by-Step Visualization of the Proposed Algorithm

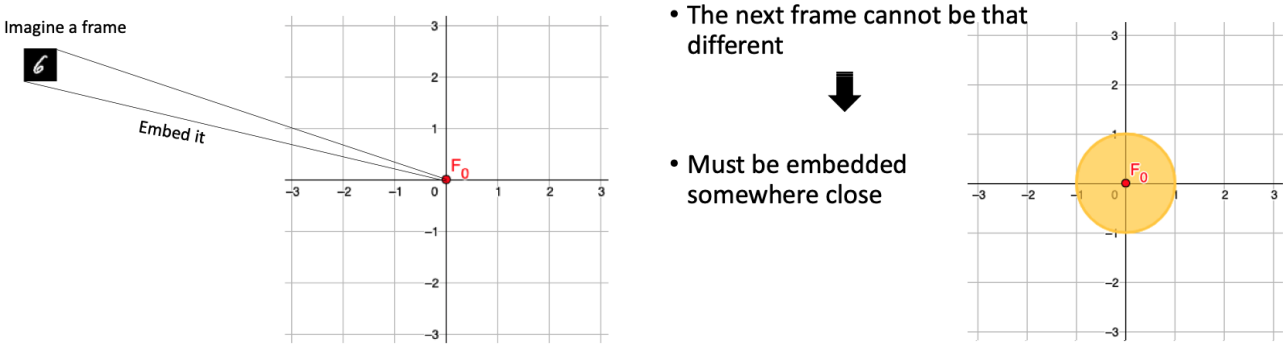
In order to help the reader further their understanding of the intuition behind our proposed algorithm we will be conducting a mental experiment with the help of some visual examples. Note that for ease of understanding the figures will follow the convention of 2+1 dimensional euclidean space.

(a) Visualization of causal intervention. Intervention is conducted by enforcing a specific path through the latent space representing the interventions in question.

(b) Visualization of causal counterfactual analysis. We represent a sample counterfactual analysis that implies different events in a prior time instant. We first identify the prior event and then take a different causal path to find our causal query $\hat{X}$ allowing us to observe it and calculate its probability of happening.

Figure 6.3: Visualization of the causal inference graphical methods.

Given a frame F_0 we embed it on our space as in Figure 6.4a. As the next frame F_1 can only have finite differences in content compared to the first, our intuition dictates that its embedding has to lie close to the original F_0, we denote this region with yellow in Figure 6.4b.

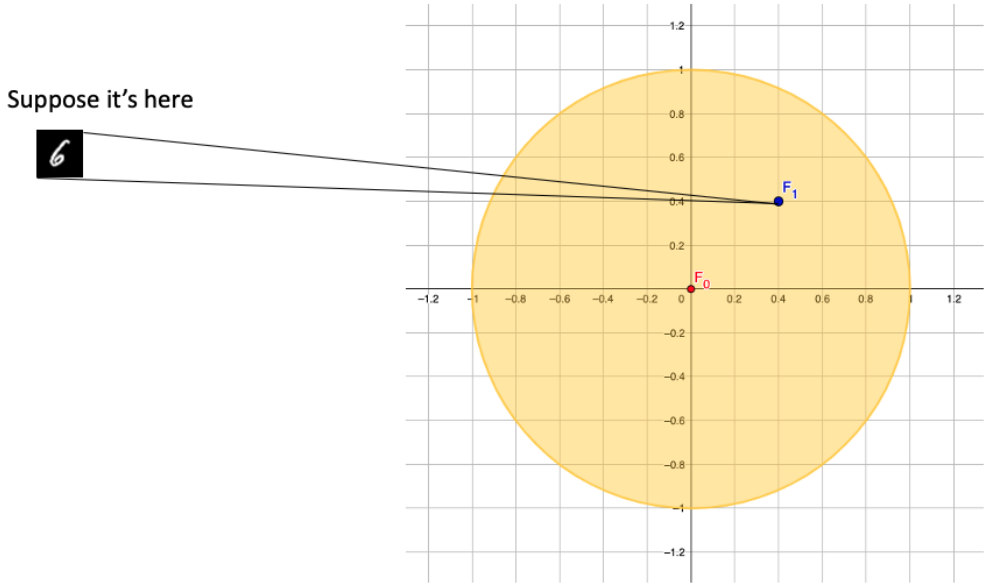


(a) Lets assume a frame F_0 that we embed in our space.

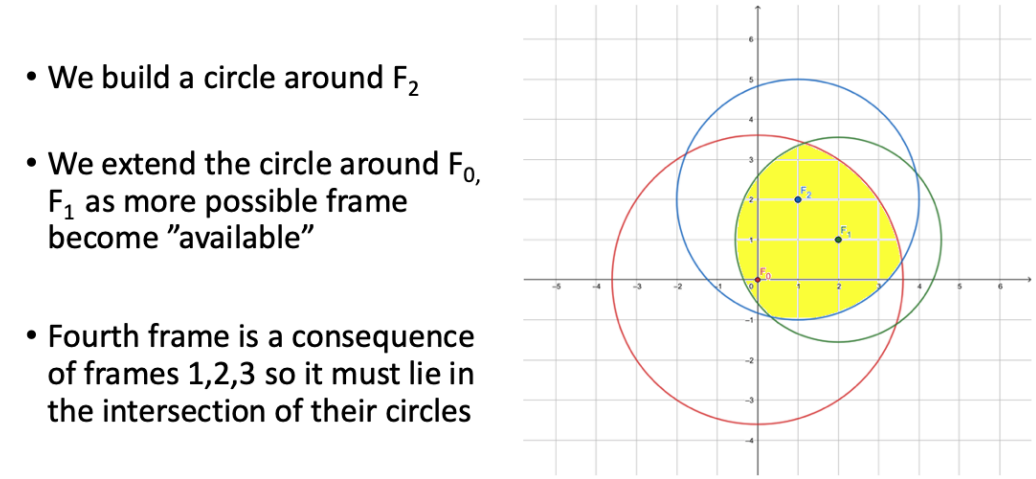
(b) The position of the next frame should be close in this space.

Figure 6.4

In Figure 6.5a we assume without loss of generality a position where F_1 will be embedded. If the next frame was known then we would simply embed it in our space in a manner similar to F_0.



(a) Embedding of a second frame.



(b) Embedding of a third frame F_2 and enlargement of the circles of frames 0,1. A fourth frame has to lie on the intersection of the circles.

Figure 6.5

Lets assume now that we have repeated the aforementioned process for a total of 3 frames and embedded them on our space. For conceptual ease, we assume frames F_0,F_1,F_2 are known a priori and we simply embed them in our space. The question that arises is where would F_3 lie? To tackle this question we have to remember that the frames constitute a sequence, hence time is also a factor that would affect our answer to the above question. In a $2D$ space we can model the passage of time by increasing the radius of the circles where the next frames lie. We base this observation on the fact that as time progresses, the content of subsequent frames can be increasingly different. Thus their embedding will be increasingly further away from our

original frame F_0.

Frame F_3, however is the consequence frame of all Frames F_0,F_1,F_2, Hence given the circles of past frames are scaled accordingly to signify the passage of time from their original time t to t' in question, the new frame has to lie in the intersection of these. As seen in figure 6.5b

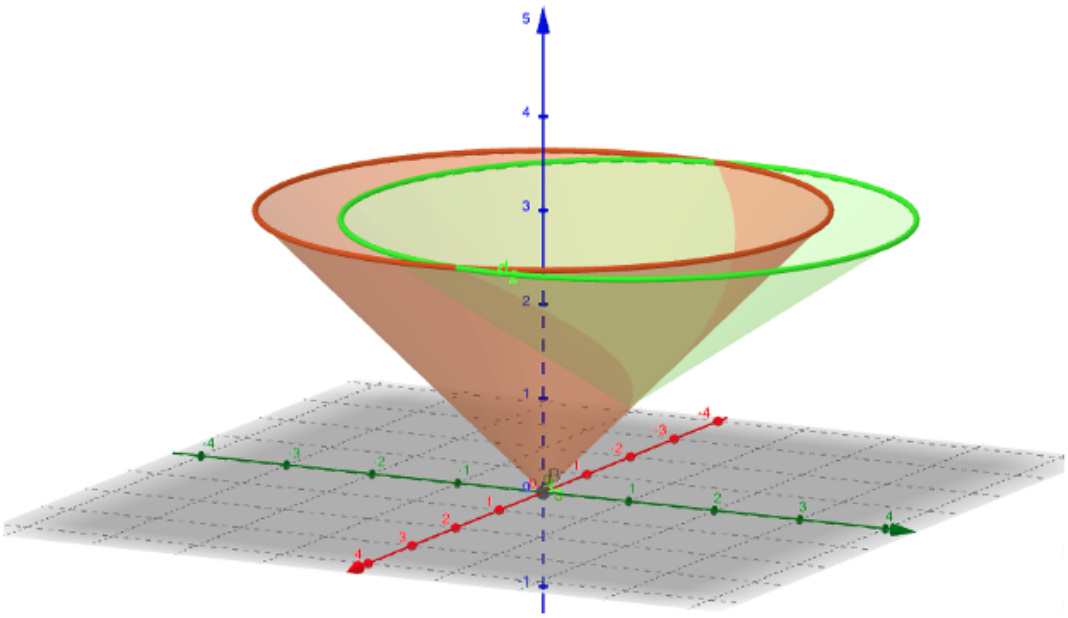


Figure 6.6: In space-time this would look like the intersection of cones. Due to the fact that the increasing radius of the 2D circles create a cone in 3D.

Our mental experiment thus far has been treating time as an invisible factor that only alters the radius of the circles. If we were to represent time as a separate observable dimension (2+1 dimensional space) then the aforementioned circles become cones, as seen in Figure 6.6. Hence the intersection of circles to find the constrained latent space where F_3 would lie becomes the intersection of cones. We visualize this in Figure 6.7

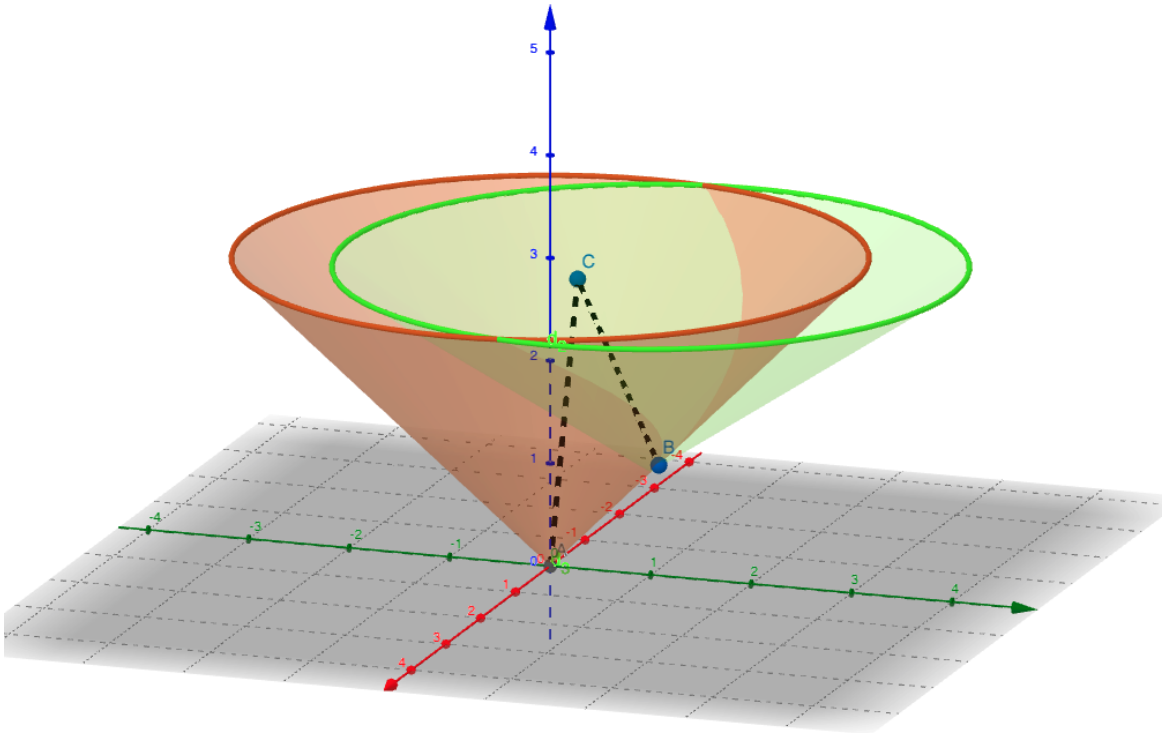


Figure 6.7: Two potential causal paths from points a,b to a new point c. Note that if its these points represent a sequence then the causal path will have to pass from $A \rightarrow B \rightarrow C$

6.4 Experimentation

6.4.1 Training

Our proposed algorithm is invariant to the method used to train the embedding. In an ideal scenario, we require an encoder-decoder pair that is able to map any image to a latent space and to reconstruct any latent code. For the purposes of this chapter's evaluation we have chosen the method by [MLM+19] as our baseline embedding, as it is the only approach that has shown good image domain performance.

[MLM+19] construct a Variational Auto Encoder (VAE) that enforces the latent space to be a Poincaré Ball. It can be shown [NK18] that a $n-$dimensional Poincaré ball embedding can be mapped into a subspace of the Minkowski space-time by an orthochronous diffeomorphism $m : P^n \rightarrow M^n$,

$$m(x_1, ...x_n) = \frac{(1 + ||x||^2, 2x_1, ..., 2x_n)}{1 - ||x||^2}, \tag{6.4}$$

and back with the inverse $m^{-1} : M^n \rightarrow P^n$

$$m^{-1}(x_1, ...x_n) = \frac{(x_1, ..., x_n)}{1 + x_0}, \tag{6.5}$$

where x_i is the i-th component of the embedding vector.

We extend [MLM+19] to enforce the embedding to a subspace of the MST by utilizing Eq. 6.4 and 6.5. We treat the space's dimensionality as hyper-parameter and tune it experimentally. We establish that the optimal embedding of our data can be achieved in an $1 + 8$ dimensional space, *i.e.*, 1 time and 8 space dimensions. The model consists of a MLP with a single hidden layer and was trained with the Riemannian equivalent of the Adam optimizer [SWKMM15] with a learning rate of $5e - 4$. Training the model with Moving MNIST requires less than one hour on a Titan RTX Nvidia GPU.

6.4.2 Inference

Our proposed Algorithm 1 is executed during inference as it does not require any learned parameters. We sample from a Gaussian distribution wrapped to be consistent with our MST in a manner similar to [MLM+19]. Inference can be performed in about 0.5 seconds per intersecting cone.

6.4.3 Dataset

As a proof-of-concept we use a custom version of the Moving MNIST dataset [SMS15]. Specifically we employ 10.000 sequences consisting of 30 frames each, making a total of 300.000 frames. Each sequence contains a single digit. The first frame is derived from the training set of the original MNIST dataset, while the subsequent frames are random continuous translations of the digit. Construction of the test set followed the same procedure with the first frame derived from the test set of the original MNIST dataset. We created 10.000 testing sequences of 30 frames each. Each frame is 32×32 while the containing digits range from $18px - 25px$.

We further use the KTH action recognition dataset by [SLC04] to highlight the real world capabilities of our method. We focus on the walking and handwaving actions and use all 4 distinct directions. Different person identities are used in the train-test split.

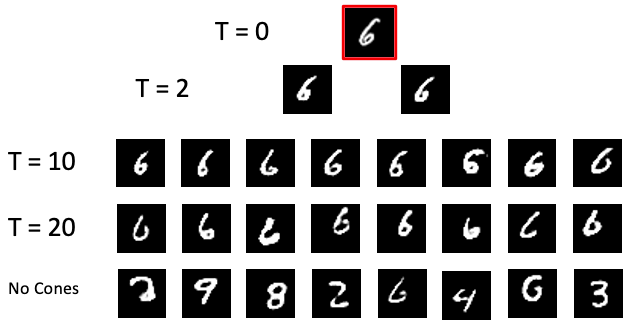
6.5 Results

6.5.1 Experiment 1: Single Cone Image Synthesis

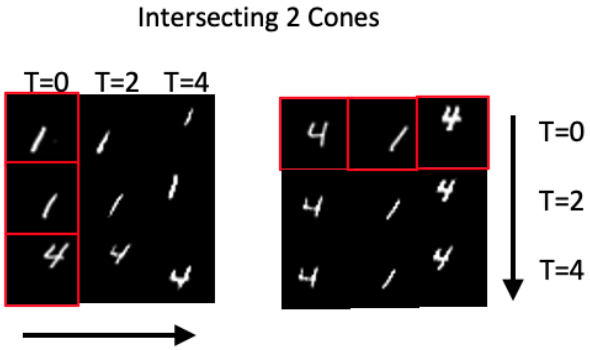
In the first experiment we evaluate the ability of the light cone to constrain the latent space such that samples lying inside the cone are reasonably similar to the original frame. We train our model with 1+8 latent dimensions. Following standard VAE sampling we produce 100.000 random samples using a wrapped normal distribution. As expected, the tighter the imposed time bound is, the fewer samples are accepted. We note that for $t = 2$ only 2 samples were accepted, for $t = 10$ our method accepts $N = 31\%$ of the samples and for $t = 20$, $N = 71\%$. In Figure 6.8a we exhibit qualitative results for Experiment 1. We note that as the time limit increases we observe higher variability, both in terms of morphology and location of the digits, while the identity of the digit remains the same. This is in accordance with the theory that the "system" would have enough time to evolve into new states.

6.5.2 Experiment 2: Intersecting Cones

In the second experiment we evaluate the ability of our algorithm to predict frames by intersecting light cones. There is no unique path a system might evolve in time. Our algorithm does not aim at producing a single future, rather it is able to produce multiple plausible futures. At a single time instant we can find any number of probable frames that extend a sequence. Hence, the `choose` step of Algorithm 1 depends on the target application. In this experiment, to guide the choice of frames, we map the sampled points to image space and compare the structural similarity of them with the original frame of $t = 0$. We adopt a simple manner to choose the next step and we do not provide the model with any further conditioning information to



(a) Samples from Experiment 1.



(b) Samples from Experiment 2.

Figure 6.8: (a): Random sampling was constrained in Experiment 1 such that the samples lie inside the light cone with an upper temporal bound. Moving from the top down the first frame $T = 0$ sets the origin of the first cone. Subsequent samples lie within the cone of the original frame with increasing time budgets cited on the left hand side. Samples in the last row of Figure (a) had no constraints imposed on them. We observe larger morphological and location differences as time progresses. This is consistent with the theory that the system had enough time to evolve into these states. (b): In Experiment 2 we are intersecting 2 cones. For ease of reading, the figures have been arranged so that the movements are more apparent. On the left in (b) we exhibit vertical movements while on the right we exhibit horizontal movements. The arrows guide the direction of reading in the figure. The red bordered frame serves as the initial seed while the $T = 2, T = 4$ frames are predicted future frames for 2 and 4 time steps respectively

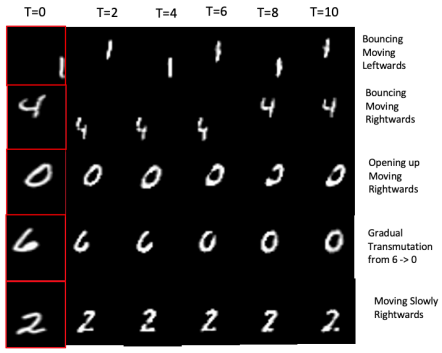
highlight the default strengths of the proposed algorithm. In an online inference scenario the reference frame could be updated as ground truth frames become available.

In Figures 6.8b and 6.9a we exhibit qualitative results of our algorithm when intersecting 2 and 5 cones respectively. In Figure 6.8b each set of results evaluates a specific movement, vertical or horizontal. In Figure 6.9a we exhibit the case of intersecting 5 cones. As this scenario allows up to 10 time steps for our model to evolve we notice a great number and more varied results. In the first two rows the depicted digits bounce while moving towards one direction. In the third row the digit 0 exhibits morphological changes and in the fourth row the digit 6 gradually moves its closing intersection upwards to become a 0. As our model is only trained with single frames of MNIST digits it is not constrained to show only movement or morphological changes. Rather it can vary both as seen in Figure 6.9a. The transmutation of the digit 6 to 0 is a probable, albeit unwanted, outcome under certain scenarios. In addition, we note that we are not providing any labels or additional information to the model during inference. In principle, one could condition the model to produce probable future frames by tuning the

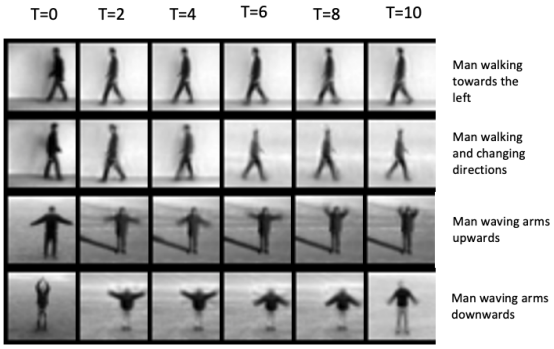
`choose` procedure of Algorithm 1.

6.5.3 Experiment 3: Realistic video data

As a third experimentation we use the KTH action dataset. Examples of the performance of the proposed algorithm are shown in Fig. 6.9b. Due to the computational constraints of the Poincaré VAE, which we are using as a base model, we are limited to one action at a time during training. We note how our algorithm retains characteristics like the shade of gray of the clothing while producing plausible futures. Each frame differs to the previous by 2 time instances giving ample time for the subject to change directions. We believe that with a higher capacity network a similar performance can be achieved on more complex scenes and higher resolution videos.



(a) Moving MNIST



(b) KTH movement video sequences dataset

Figure 6.9: Samples from Experiment 2 (a) and Experiment 3 (b). We are intersecting 5 cones trained on the moving MNIST dataset in (a) and the KTH movement video sequence dataset in (b). The latter is representative for a real-world use-case scenario. Next to each row we added an explanatory caption about the type of observed movement. Differences in image brightness in (b) are due to PyTorch's contrast normalization in the plotting function.

6.5.4 Experiment 4: Prediction drifting

We evalute the accumulation of autoregressive errors for the task of future prediction, as well as indications about possible causal breaches. As we mentioned before, autoregressive solutions to the prediction problems incur accumulation of errors that lead to degradation of the visual quality of the produced images. In methods that perform prediction tasks in the latent space

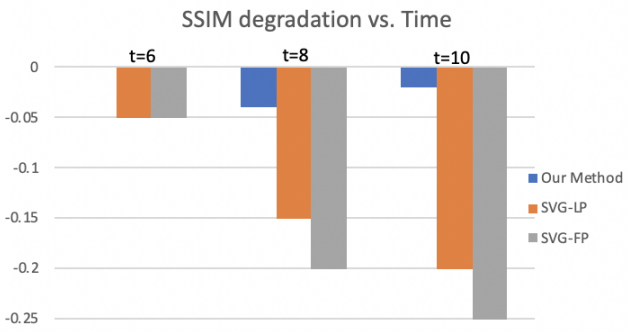
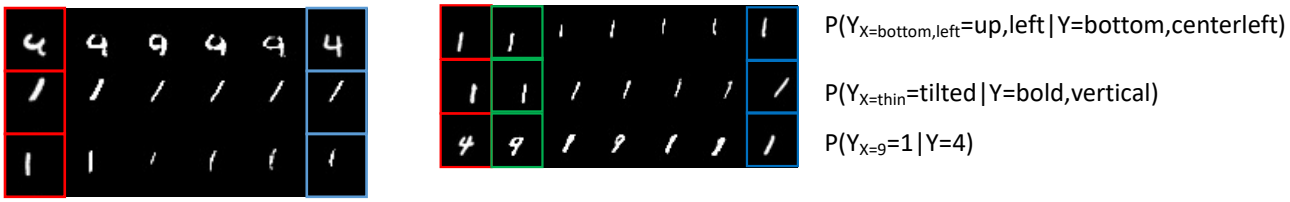


Figure 6.10: Structural Similarity Metric (SSIM) comparison of our method against [DF18]. In time instant 6 our method produces effectively 0 ssim error and remains low up to 10 time steps into the future, while both fixed prior and learned prior SVG methods degrade rapidly.

such errors decrease. In addition, other methods do not have guarantees of causal performance, leading to potential morphological changes of the object in the image or non-continuous motions. In order to evaluate the above two challenges in our method we assess the Structural Similarity Metric (SSIM) of the produced frames up to 10 time steps into the future. We compare against a well known method [DF18] that uses the moving MNIST dataset for future prediction. In Figure 6.10 we show the results of the average of 100 runs of the SSIM error of our method and the fixed and learned prior versions of [DF18]. We compare the generated frames against their ground truth sequence frames. For the *choose* step of our algorithm, we randomly pick a generated frame that lies inside the lightcone of the input frame. We note that while [DF18] degrades rapidly as time progresses, our method does not. Indeed, on average our method provides negligible SSIM error up to 6 time steps into the future. We further note that our method only uses a single input frame in this experiment compared to the two or more required by the SVG method.

We hypothesize that a significant drift in the SSIM measure could be the effect of either non-causal drifting, *i.e.*, the method producing frames that are not causally related to the original frames, due to autoregressive error compounding, or a combination of the two. Trivial solutions of just outputting the same frame would, indeed, lead to a negligible SSIM error. However, as seen in Figures 6.8a, 6.8b and 6.9a our method does not collapse to trivial outputs. Hence, we can argue that the effects of autoregressive error compounding are minimized together with errors from generated non-causally linked frames.



(a) Intervention (b) Counterfactual

Figure 6.11: Causal Inference examples for both interventions and counterfactuals, Red is the original conditioning frame; Blue is the target frame to reach with interventions and Green is the counterfactual conditioning frame.

6.5.5 Experiment 5: Causal Inference

In the final experiment we explore the aforementioned path-based causal inference processes. In Figure 6.11 we show examples for both, interventions and counterfactuals. In both cases to guide interventions we give the model a target frame depicted in blue. Subsequent frames were chosen after random sampling given that their position in the latent space approached the target frame. We observe that in both cases our algorithm is able to generate a plausible path from the initial query frames to the target ones. In such a way one could explore the potential outcomes and possible paths to a query state, hence, extracting causal intuition about the problem at hand. In relation to counterfactuals we do not generate a single counterfactual, but rather a series of possible counterfactuals, as we do not impose any identifiability monotoniciy constraints, as detailed in [Pea09]. Contrary to other synthesis methods that abide by the Abduction-Action-Prediction or the Twin network paradigm this method, does not make any strong assumptions on the underlying causal diagram or suffers from the computational scaling laws of Abduction-Action-Prediction. Moreover as we produce a series of potential outcomes the "select" step of our algorithm is crucial. For simplicity we chose the next step randomly but one could envision specific selection algorithms that allow the incorporation of domain knowledge or application specific filtering. These additions, however, do come with the trade-off that an increase complexity of the selection step exponentially increases the algorithms complexity as this step is repeated after each sampling.

6.6 Discussion

As the model is only trained as a VAE on single frames and not on sequences, the notion of time is not encoded in the weights of the network. Hence, all the resulting movement and predictive abilities are derived from our proposed algorithm and the natural embedding abilities of the MST.

We emphasize the time-agnostic nature of our algorithm. With this we denote that our ML model does not learn the concept of time, but of all possible scenarios. Time is incorporated by the way we parse and constrain the latent space. Our predictions are constrained in time but are probabilistic in nature. The proposed algorithm is able to produce multiple plausible futures. We believe this is a very important feature for future prediction and sequence extrapolation techniques as it can be used as an anomaly detection technique. Specifically, if one of the produced futures includes a hazardous situation, an automated system can adapt in order to avoid an outcome, enabling for example defensive driving capabilities in autonomous vehicles.

Even though our method is in principle auto-regressive, it does not suffer from the accumulation of errors as it is both probabilistic and relies on efficient latent space sampling rather than the ability of a neural network to remember structural and temporal information about the image.

Furthermore, we believe that the quality of the predicted frames as well as the definition of the subspace from which the samples should be derived could be improved by incorporating the inferred thermodynamic entropy of the frame. We will explore the link between the information and thermodynamic entropy in future work. In addition, even though our framework is architecture agnostic, a customized architecture for the prediction task would be an intriguing direction. We believe that the use of the Poincaré VAE heavily reduces our abilities to produce good visual quality samples due to its extreme difficulty to train. Potential future development avenues could include the use of diffusion models instead.

Finally, as our model allows us to find all probable scenarios that might exist, it can be used as a causal inference tool in the “potential outcomes” framework by [Rub05]. Given a state we are able to probe possible scenarios and investigate plausible outcomes, hence, deduce causal

relations within the data. In addition, by using the *past* light cone $\mathcal{R}^-$, we are able to probe the events that could have led to an observed state enabling counterfactual analysis.

6.7 Summary

Machine learning techniques are able to build powerful representations of large quantities of data. We leverage this ability and propose that hard computer vision problems can be approached with minimal learning in an architecture agnostic manner. Strong mathematical and physical priors are the key. In this chapter, we extend early Riemannian representation learning methods with the notion of Minkowski space-time as it is more suitable for causal inference. We further propose a novel algorithm to perform causally plausible image synthesis and future video frame prediction utilizing the special relativity concept of light cones. We showed our algorithms' capabilities both in the synthetic Moving MNIST dataset as well as the real-world KTH dataset.

Chapter 7

Counterfactual Inference in Tabular and Medical Data

In the previous chapter we discussed how to predict potential future directions of videos in a causally enabled way. In this chapter we will dive a bit deeper into the causal analysis of phenomena and establish methods to assess counterfactual queries. Figure 7.1 shows the positioning of the ability of answering counterfactual questions in relation with the other modules of an ADS. Even though our previous approach was computationally lighter than other works in the field, we still abode by the common *Abduction-Action-Prediction* paradigm. In this chapter we explore an alternative paradigm - the twin networks. We first evaluate this novel approach in a wide array of medical and financial data. More crucially, though, in this chapter we address one of the major requirements of our main application. How do we ensure our causal predictions are accurate, and we have learned to correct model. This is a topic commonly known as *identifiability*. This chapter is based upon the associated publication at the causal workshops of UAI 2022 and ICML 2022, while the full paper is under review at Nature Machine Intelligence [VKGL21].

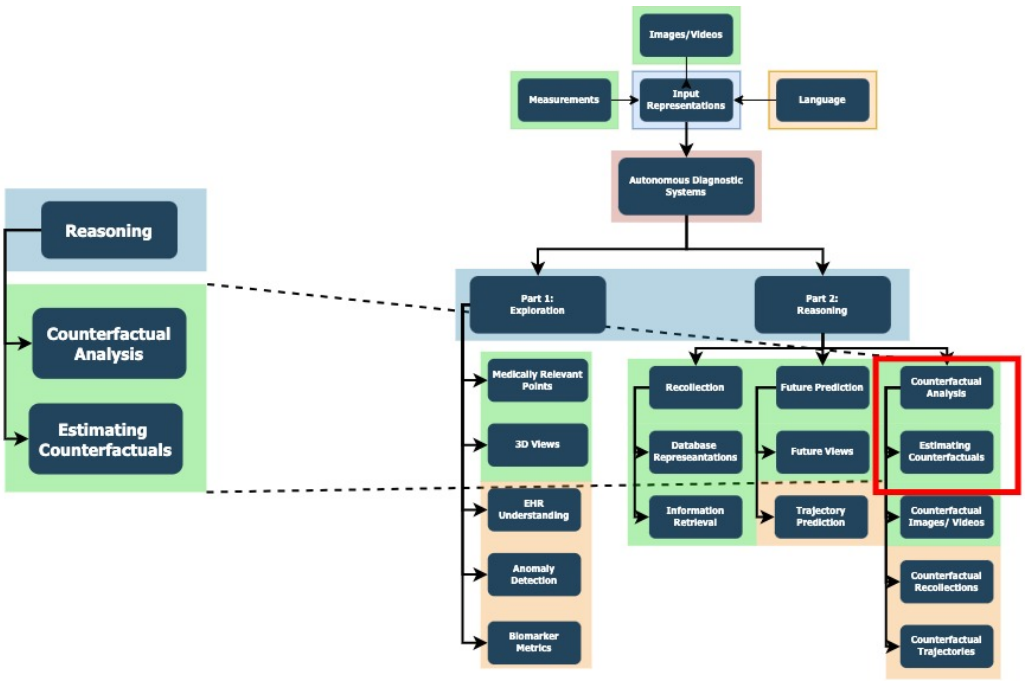


Figure 7.1: In this chapter we look into how to assess counterfactual queries

7.1 Introduction

"If my credit score had been better, would I have been approved for this loan?", "What is the effect of the diabetes type on the risk of stroke?". Causal questions like these are routinely asked by scientists and the public alike. Recent machine learning advances have enabled the field to address causal questions in high-dimensional datasets to a certain extent [SLK18, AWVDS17, SBV19]. However, most of these methods focus on `Interventions`, which only constitute the second-level of Pearl's three-level causal hierarchy [Pea09, BCII20]. At the top of the hierarchy sit `Counterfactuals`. These subsume interventions and allow one to assign fully causal explanations to data.

Counterfactuals investigate alternative outcomes had some pre-conditions been different. The crucial difference between counterfactuals and interventions is that the evidence the counterfactual is "counter-to" can contain the variables we wish to intervene on or predict. The first question posed at the start of this paper, for instance, is a counterfactual one. Here we want to know if improving our credit score will lead to loan approval in the explicit context that the loan has just been declined. A corresponding interventional query would be "what is the impact of the credit score on the loan approval chances?". Here, evidence that the loan has just been

denied is *not* used in estimating the impact. The second question posed at the start of this paper—regarding the effect of diabetes type on the risk of stroke—is an interventional question. By utilizing this additional information, counterfactuals enable more nuanced and personalized reasoning and decision making. Counterfactual inference has been applied in high profile sectors like medicine [RLJ20, OS19], legal analysis [LGZ13], fairness [KLRS17], explainability [GPS21], and advertising [AL19].

To perform counterfactual inference, one requires knowledge of the causal mechanisms. However, the causal mechanisms cannot be uniquely determined from observations and interventions alone. Indeed, two causal models that have the same conditional and interventional distributions can disagree about certain counterfactuals [Pea09]. Hence, without additional constraints on the form of the causal mechanisms, they can generate "non-intuitive" counterfactuals that conflict with domain knowledge, as originally pointed out by [OS19].

This raises the question of how best to choose the causal mechanisms so that resulting counterfactual inference is trustworthy in a given domain. Despite the importance of counterfactual inference, this question has only been addressed in causal models with binary treatment and outcome variables [TP00]. The case of categorical variables remains unanswered. Beyond binary variables, previous work has only derived upper and lower bounds for counterfactual probabilities [ZB20]. In many cases, these bounds can be too wide to be informative. We address this challenge by introducing for causal models with categorical variables the notion of *counterfactual ordering*, a principle that posits desirable properties causal mechanisms should possess, and prove that it is equivalent to specific functional constraints on the causal mechanisms. Namely, we prove that causal mechanisms satisfying counterfactual ordering must be monotonic functions.

To learn such causal mechanisms, and perform counterfactual inference with them, we introduce *deep twin networks*. These are deep neural networks that, when trained, are capable of *twin network* counterfactual inference—an alternative to the *abduction, action, & prediction* method of counterfactual inference. Twin networks were introduced by [BP94] and reduce estimating counterfactuals to performing Bayesian inference on a larger causal model, known as a

twin network, where the factual and counterfactual worlds are jointly graphically represented. Despite their potential importance, twin networks have not been widely investigated from a machine learning perspective. We show that the graphical nature of twin networks makes them particularly amenable to deep learning.

We empirically test our approach on a variety of real and semi-synthetic datasets from medicine and finance, showing our method achieves accurate estimation of counterfactual probabilities. Moreover, we demonstrate that if counterfactual ordering is not enforced, the model generates "non-intuitive" counterfactuals that contradict domain knowledge in these cases. Our contributions are as follows:

1. We introduce *counterfactual ordering* for causal models with categorical variables, which posits desirable properties causal mechanisms should posses.
2. We prove *counterfactual ordering* is equivalent to specific functional constraints on the causal mechanisms. Namely, that they must be monotonic.
3. We introduce *deep twin networks* to learn such causal mechanisms and perform counterfactual inference. These are deep neural networks that, when trained, can perform *twin network* counterfactual inference.
4. We test our approach on real and semi-synthetic data, achieving *accurate* counterfactual estimation that complies with domain knowledge.

7.2 Background

We invite the reader to refer back to Section 2.9 for an in-depth view on counterfactual inference. In this chapter we will be expanding the notions covered in the beginning of this thesis to counterfactual inference using the twin networks and the ability to identify causal effects from observational data.

7.2.1 Twin network counterfactual inference

A practical limitation of Theorem 2.1 is that the abduction step requires large computational resources. Indeed, even if we start with a Markovian model in which background variables are mutually independent, conditioning on evidence—as in abduction—normally destroys this independence and makes it necessary to carry over a full description of the joint distribution over the background variables [Pea09]. [BP94] introduced a method to address this difficulty. Their method reduces estimating counterfactuals to performing Bayesian inference on an larger causal model, known as a *twin network*, where the factual and counterfactual worlds are jointly graphically represented, described in technical detail below.

As was discussed above, abduction-action-prediction counterfactual inference requires a lot of computational resources. Twin networks were specifically designed to address this difficulty [Pea09, BP94]. Indeed, consider the following passage from Pearl's "Causality" [Pea09, Section 7.1.4, page 214]:

> "The advantages of delegating this computation [abduction] to inference in a Bayesian network [i.e., a twin network] are that the distribution need not be explicated, conditional independencies can be exploited, and local computation methods can be employed"

This suggests that the computational resources required for counterfactual inference using twin networks can be less than in abduction-action-prediction. This was put to the test by [GLP19] and shown empirically to be correct, with their abstract stating

> "twin networks are faster and less memory intensive by orders of magnitude than standard [abduction-action-prediction] counterfactual inference"

A key difference is that in a twin network, inference can be conducted in parallel rather than in the serial nature of abduction-action-prediction. For instance, sampling in twin networks is faster than in the abduction-action-prediction, as twin networks propagate samples simultaneously through the factual and counterfactual graphs—rather than needing to update, store

and resample as in abduction-action-prediction. Thus, full counterfactual inference in a twin network can take up to no more than the amount of time sampling takes, while in abduction-action-prediction one incurs the additional cost of reusing samples and evaluating function values in the new mutilated graph. This is potentially advantageous for very large graphs, or for graphs with complex latent distributions that are expensive to sample.

We now remind the reader the details of the twin network. A twin network consists of two interlinked networks, one representing the real world and the other the counterfactual world being queried. Constructing a twin network given a structural causal model and using it to compute a counterfactual query is as follows: First, one duplicates the given causal model, denoting nodes in the duplicated model via superscript *. Let $V = \{v_1, \ldots, v_n\}$ be observable nodes in the causal model and $V^* = \{v_1^*, \ldots, v_n^*\}$ the duplication of these. Then, for every node v_i^* in the duplicated, or "counterfactual," model, its latent parent u_i^* is replaced with the original latent parent u_i in the original, or "factual," model, such that the original latent variables are now a parent of two nodes, v_i and v_i^*. The two graphs are linked only by common latent parents, but share the same node structure and generating mechanisms. To compute a general counterfactual query $P(Y = y \mid E = e, \mathrm{do}(X = x))$, one modifies the structure of the counterfactual network by dropping arrows from parents of X^* and setting them to value $X^* = x$. Then, in the twin network with this modified structure, one computes the following probability $P(Y^* = y \mid E = e, X^* = x)$ via standard inference techniques, where E are factual nodes. That is, in a twin network one has:

$$\begin{aligned} P(Y = y \,|E = e, \mathrm{do}(X = x)) = \\ P(Y^* = y \mid E = e, X^* = x) \end{aligned} \tag{7.1}$$

To illustrate this concretely, consider the causal model with causal structure depicted in Figure 2.8a, where variables X, Y are binary. The counterfactual statement to be computed is $P\left(Y = 0 \mid Y = 1,\ \mathrm{do}(X = 0)\right)$. The twin network approach to this problem first constructs the linked factual and counterfactual networks depicted in 2.8b. The intervention $\mathrm{do}(X^* = 0)$ is then performed in the counterfactual network; all arrows from the parents of X^* are removed

and X^* is set to the value 0—graphically depicted in 2.8c. The above counterfactual query is reduced to the following conditional probability in 2.8c: $P\left(Y^* = 0 \mid Y = 1, X^* = 0\right)$, which can be computed using Bayesian inference techniques.

Our contribution: Despite their importance for counterfactual inference, twin networks have not been widely studied—particularly from a machine learning perspective. In the Methods section we demonstrate how to combine twin networks with neural networks to estimate counterfactuals.

7.2.2 Non-identifiability of counterfactuals

As a general principle, observational and interventional data do not allow for identification of counterfactual distributions, as multiple parametrizations of the causal phenomenon can be consistent with the observed data. Eq. 7.2 illustrates this phenomenon with a simple example. With our example we aim to emphasise how in such a simple case where identifiability is violated all models agree on the observational and interventional quantities but disagree on the counterfactuals. As such without extra information we are not able to determine which is the correct model. Lets assume the exemplary case of DAG 2.8e, with X, Y *binary*, U_Y a four-valued variable distributed under $q(U_Y)$ and $\neg X$ to be the logical negation of X. U_Y and its distribution $q(U_Y)$ is a latent noise variable that represents the probabilistic element in our otherwise deterministic DAG .

$$Y = \begin{cases} X, & \text{if } U_Y = 0 \\ 0, & \text{if } U_Y = 1 \\ 1, & \text{if } U_Y = 2 \\ \neg X, & \text{if } U_Y = 3 \end{cases} \tag{7.2}$$

In this case the conditional probabilities are given by $P(Y \mid X)$ are: $P(Y = 0|X = 0) = q(U_Y = 0) + q(U_Y = 1)$ and $P(Y = 0|X = 1) = q(U_Y = 1) + q(U_Y = 3)$. As there are no confounders

in our example, these coincide with the interventional distributions $P(Y \mid \mathrm{do}(X))$.

The causal model in Eq. 7.2 has 3 parameters, but the conditional distributions only provide 2 constraints on these parameters. Hence, due to the existence of this free parameter, there can exist models with the same conditional distributions, but different counterfactuals. This observation can bee seen as follows. Consider the following counterfactual query $P(Y_{X=1} = 1 \mid Y = 0, X = 0)$. It can be written as:

$$P(Y_{X=1} = 0 \mid Y = 1, X = 0) = \frac{q(U_Y = 3)}{q(U_Y = 2) + q(U_Y = 3)}$$

which follows by noting $Y = 1$ and $X = 0$ implies either $Y = \neg X$ or $Y = 1$, which happens with probability $q(U_Y = 2) + q(U_Y = 3)$, and within this context the only way $Y = 0$ when $X = 1$ is if $Y = \neg X$, which occurs with probability $q(U_Y = 3)$. Note that there is no way to write this counterfactual probability in terms of the conditional distributions $P(Y \mid X)$ alone. Moreover, one can have two models with the same functional form as Eq. 7.2 that yield the same conditional distributions, but give *different* predictions for the above counterfactual query. An example is one model with distribution over U_Y given by $\{q(U_Y)\}_0^3 = \{1/2, 1/6, 1/6, 1/6\}$, another with $\{q(U_Y)\}_0^3 = \{1/3, 1/3, 1/3, 0\}$. Hence, there are counterfactuals that are not *identifiable* from data. Such counterfactual distributions cannot be expressed in terms of observational or interventional distributions alone. This can lead to counterfactual predictions that conflict with domain knowledge, as shall be shown in the next section.

7.3 Methods

7.3.1 Non-identifiability & domain knowledge

Is non-identifiability of counterfactuals a problem? Given a causal model trained on observations and interventions, can we always trust its counterfactual predictions? In general the answer is no: counterfactual predictions from a causal model can conflict with domain

knowledge—even if it perfectly reproduces observations and interventions, as we now show.

In epidemiology, causal models with the structure of Figure 2.8e are studied, where X is the presence of a risk factor and Y is the presence of a disease. From epidemiological domain knowledge, it is believed that risk factors always increase the likelihood of a disease being present [TP00]—referred to as "no-prevention", that no individual in the population can be helped by exposure to the risk factor [Pea99]. Hence, if one observes a disease, but not the risk factor, then, in that context, if we had intervened to give that individual the risk factor, the likelihood of them not having the disease must be zero—as having the risk factor can only increase the likelihood of a disease.

We'll now describe two causal models that generate the same observations and interventions, yet the counterfactuals generated by one model satisfy the above domain knowledge and the other do not. Consider the two different parameterizations of the causal model discussed in Section 7.2.2: $\{q(U_Y)\}_0^3 = \{1/2, 1/6, 1/6, 1/6\}$ and $\{q(U_Y)\}_0^3 = \{1/3, 1/3, 1/3, 0\}$. Both models have the same conditional distributions, and we have $P(Y_{X=1} = 1) > P(Y_{X=0} = 1)$ and $P(Y_{X=1} = 0) < P(Y_{X=0} = 0)$. This tells us that intervening to set $X = 1$ *always* makes $Y = 1$ more likely, and doesn't increase the likelihood of $Y = 0$ relative to $X = 0$. So, at the interventional-level, these models seem to comply with Epidemiological domain knowledge that says that the presence of a risk factor, that is, $X = 1$, always makes disease, $Y = 1$, more likely.

Despite this, in the first model $P(Y_{X=1} = 1 \mid Y = 1, X = 0) < P(Y_{X=0} = 1 \mid Y = 1, X = 0)$. According to this model, $Y = 1$ becomes *less* likely when we intervene with $X = 1$ in the counterfactual context $Y_{X=0} = 1$—even though intervening to set $X = 1$ can only make $Y = 1$ more likely, and does not increase the likelihood of $Y = 0$. This is a very "non-intuitive" counterfactual prediction from the point of view of an Epidemiologist. However, in the second model, $P(Y_{X=1} = 1 \mid Y = 1, X = 0) = P(Y_{X=0} = 1 \mid Y = 1, X = 0)$. Indeed, in this model, no matter the counterfactual context, intervening to set $X = 1$ *never* reduces the likelihood $Y = 1$. Thus the second model complies fully with Epidemiological domain knowledge.

As the models agree on the data they're trained on, we must impose extra constraints to learn the model that generates domain-trustworthy counterfactuals. In the next section, we present

a simple principle that provide such constraints.

7.3.2 Counterfactual ordering

We continue to consider the causal structure from Section 7.2.2 with a DAG as depicted in Figure 2.8e. However, now X, Y are categorical variables with an arbitrary number of categories N, M each. Inspired by the Epidemiological example from the previous section, we now define *counterfactual ordering*, which posits an intuitive relationship between counterfactual and intervantional distributions.

Definition 7.1 (Counterfactual Ordering)**.** *A causal model with categorical treatment variable X and categorical outcome variable Y satisfies* counterfactual ordering *if there exists an ordering on interventions and outcomes $\{x_0, x_1, \ldots, x_N\}$, $\{y_0, y_1, \ldots, y_M\}$ such that $P(Y_{x_i} = y_k) \geq P(Y_{x_j} = y_k)$ and $P(Y_{x_i} = y_h) \leq P(Y_{x_j} = y_h)$ for all $i > j$ and $k > h$, then it must be the case that $P(Y_{x_i} \geq y_k | Y_{x^*} = y^*) \geq P(Y_{x_j} \geq y_k | Y_{x^*} = y^*)$ for all counterfactual contexts $\{y^*, x^*\}$.*

This encodes to the following intuition: If intervention x_i only increases the likelihood of outcome y_k, relative to any intervention x_j with $j < i$, without increasing the likelihood of y_h for all $h < k$, then intervention x_i must increase the likelihood that the outcome we observe is at least as high as y_k, regardless of the context. Counterfactual ordering places the following constraints on a causal model.

Theorem 7.1. *If counterfactual ordering holds $P(Y_{x_j} = y_l | Y_{x_i} = y_h) = 0$ for all $l > h$ and $i > j$.*

Proof. First note that $P(Y_{x'} = y' | Y_{x'} = y) = 0$ for any $y' \neq y$ follows from the definition of counterfactuals. The conjunction of this and counterfactual ordering implies $0 = P(Y_{x_i} \geq y_k | Y_{x_i} = y_h) \geq P(Y_{x_j} \geq y_k | Y_{x_i} = y_h)$. As probabilities are bounded below by 0, we have $P(Y_{x_j} \geq y_k | Y_{x_i} = y_h) = \sum_{l>h} P(Y_{x_j} = y_l | Y_{x_i} = y_h) = 0$. Again, as probabilities are non-negative, we have $P(Y_{x_j} = y_l | Y_{x_i} = y_h) = 0$ for all $l > h$ and $i > j$. □

Equalities of the form $P(Y_x = y'|Y_{x'} = y) = 0$ are equivalent to the statement $\{Y_x = y'\} \wedge \{Y_{x'} = y\}$ = False, where $\wedge$ is the logical AND operator. Therefore, the conjunction of the input-output pairs $X = x, Y = y'$ and $X = x, Y = y$ cannot occur in such a causal model. This yields constraints on the model parameters beyond those imposed by observations and interventions that can be enforced during causal model training.

It is important to note that we are not saying every causal model should satisfy counterfactual ordering. As in all works on causal inference, it is ultimately up to the analyst to decide if such an assumption appears reasonable in a given domain. As discussed in the previous section, counterfactual ordering appars a reasonable assumption in Epidemiology. In the Experiments section, we empirically demonstrate on data from medicine and finance that models that are trained to satisfy counterfactual ordering comply with domain knowledge, while models that aren't appear in conflict with domain knowledge.

7.3.3 Counterfactual ordering and Counterfactual Stability

Counterfactual stability has been proposed by [OS19] as a different way to restrict the type of counterfactuals a causal model can output, to ensure they are "intutive". We define counterfactual stability then prove a relation between it and counterfactual ordering.

Definition 7.2 (Counterfactual Stability)**.** *A causal model of categorical variable Y satisfies* counterfactual stability *if it has the following property: If we observe $Y_x = y$, then for all $y' \neq y$, the condition $\frac{P(Y_x=y)}{P(Y_{x'}=y')} \geq \frac{P(Y_x=y')}{P(Y_{x'}=y)}$ implies that $P(Y_x = y'|Y_{x'} = y) = 0$. That is, if we observed $Y = y$ under intervention $X = x$, then the counterfactual outcome under intervention $X = x'$ cannot be equal to $Y = y'$ unless the multiplicative change in $P(Y_x = y)$ is less than the multiplicative change in $P(Y_x = y')$.*

This encodes the following intuition about counterfactuals: If we had taken an alternative action that would have only increased the probability of $Y = x$, without increasing the likelihood of other outcomes, then the same outcome would have occurred in the counterfactual case. Moreover, in order for the outcome to be different under the counterfactual distribution, the

relative likelihood of an alternative outcome must have increased relative to that of the observed outcome.

Counterfactual stability is weaker than counterfactual ordering, as it imposes fewer constraints on the model. In fact, counterfactual ordering between intervention and outcome values implies counterfactual stability holds between them:

Theorem 7.2. *If a causal model satisfies* counterfactual ordering *then it satisfies* counterfactual stability.

Proof. First, Theorem 7.1 says counterfactual ordering implies $P(Y_x = y'|Y_{x'} = y) = 0$. Additionally, we need to show that when counterfactual ordering holds, $\frac{P(Y_x=y)}{P(Y_{x'}=y')} \geq \frac{P(Y_x=y')}{P(Y_{x'}=y)} \implies P(Y_x = y'|Y_{x'} = y) = 0$. From counterfactual ordering we have $P(Y_x = y) \geq P(Y_{x'} = y)$ and $P(Y_x = y') \leq P(Y_{x'} = y')$. The latter implies $\frac{P(Y_x=y')}{P(Y_{x'}=y')} \leq 1$, which when combined with the former yields: $P(Y_x = y) \geq \frac{P(Y_x=y')}{P(Y_{x'}=y')} P(Y_{x'} = y)$.. Concluding the proof. □

7.3.4 Counterfactual ordering functionally constrains causal mechanisms

[OS19] were unable to derive any general functional constraints counterfactual stability places on the causal mechanisms underlying a given causal model. They were only able to compute counterfactuals satisfying it in a single, specific type of causal model. Namely, one where the mechanisms are parameterised using the Gumbel-Max trick. By contrast, we now derive general a functional constraint on causal mechanisms that is equivalent to counterfactual ordering. In the next section we will show how to learn causal models that satisfy this constraint. Thus we are able to learn causal models that satisfy counterfactual ordering without the need for specific parametric assumptions—such as the Gumbel-Max trick, as was required for counterfactual stability by [OS19].

Definition 7.3 (Monotonicity)**.** *If there exists an ordering on interventions and outcomes:* $\{x_0, x_1, \ldots, x_N\}$, $\{y_0, y_1, \ldots, y_M\}$ *such that* $P(Y_{x_i} = y_k) \geq P(Y_{x_j} = y_k)$ *and* $P(Y_{x_i} = y_h) \leq$

$P(Y_{x_j} = y_h)$ for all $i > j$ and $k > h$ then $Y_x(u) \geq Y_{x'}(u)$ for all u. Equivalently, the events $\{Y_{x_i} = y_h\} \wedge \{Y_{x_j} = y_l\} =$ False for all $i > j$ and $h < l$.

Theorem 7.3. *Given an intervention & outcome ordering, counterfactual ordering & monotonicity are equivalent.*

Proof. First we show counterfactual ordering implies monotonicity. From Theorem 7.1, counterfactual ordering implies $\{Y_{x_i} = y_h\} \wedge \{Y_{x_j} = y_l\} =$ False for all $i > j$ and $h < l$, as $P(Y_{x_i} = y_h | Y_{x_j} = y_l) = 0$. Monotonicity follows.

Next we show monotonicity implies counterfactual ordering. Monotonicity implies that for intervention $X = x$, the likelihood of the outcome being higher in the outcome ordering increases, while the likelihood of the outcome being lower in the ordering decreases relative to the likelihoods imposed by intervention $X = x'$ which lies lower in the intervention ordering. All that remains is to show that for such interventions, x, x', and outcome $Y = y$ for which $P(Y_x = y) \geq P(Y_{x'} = y)$, it follows that $P(Y_x \geq y | Y_{x^*} = y^*) \geq P(Y_{x'} \geq y | Y_{x^*} = y^*)$ for all counterfactual contexts $\{y^*, x^*\}$.

$P(Y_x \geq y | Y_{x^*} = y^*)$ is computed by first updating $P(U)$ under x^*, y^* and computing $P(Y \geq y)$ in the submodel M_x. That is, it corresponds to the expected value of $P(Y_x(u) \geq y)$ under $u \sim P(U | x^*, y^*)$. From monotonicity one has $Y_x(u) \geq Y_{x'}(u)$ for all u. Hence, for any $U = u$ that results in $Y_{x'}(u) \geq y$, that same $U = u$ yields $Y_x(u) \geq y$, as $Y_x(u) \geq Y_{x'}(u) \geq y$. Hence, as there are at most as many values of U that lead to $Y_x \geq y$ as lead to $Y_{x'} \geq y$, one has $P(Y_x(u) \geq y) \geq P(Y_{x'}(u) \geq y)$ for any $U = u$. This follows because $P(Y_x(u) \geq y) = \sum_{u | Y_x(u) \geq y} P(U = u)$ together with the observation that summands in $\sum_{u | Y_{x'}(u) \geq y} P(U = u)$ are a subset of summands in $\sum_{u | Y_x(u) \geq y} P(U = u)$. Taking expectations under $P(U | y^*, x^*)$ yields the proof. □

[TP00] proved that in an SCM with DAG 2.8a with binary X, Y, where Y is monotonic in X, the probabilities of causation—important counterfactual queries that quantify the degree to which one event was a necessary or sufficient cause of another—can be uniquely identified from observational and interventional distributions. See Section 7.3.6 for a definition of the probabilities of causation. We thus have the follow corollary to theorem 7.3.

Corollary 1. *In a counterfactually ordered SCM with DAG 2.8a and binary X, Y, the probabilities of causation are identified from observational and interventional distributions.*

For categorical variables beyond the binary case, it is unknown whether monontonicity implies unique identifiability. However, in this work we are not concerned with counterfactuals being uniquely defined, as long as "non-intuitive" counterfactuals are ruled out. Constraints on the model beyond those imposed by observations and experimental data are said to *partially-identify* counterfactual distributions. In the next section we demonstrate how to learn causal models satisfying counterfactual ordering from data.

7.3.5 Deep twin networks

We now present *deep twin networks* which combine twin networks with neural networks to learn the causal mechanisms and estimate counterfactuals. Importantly, we will discuss how to ensure the function space learned by the neural network satisfies counterfactual ordering.

Contrary to prior Bayesian network approaches deep twin networks allow us not only to estimate counterfactual probabilities from data but to learn the underlying functions that dictate the interactions between causal variables. Hence we are able to gain a deeper insight on the mechanisms represented in the structural causal model that generate the counterfactuals we want to predict. Moreover we are able to quantify the uncertainty about the outcome by learning the latent noise distribution. Finally, the use of neural networks allow us to gain flexibility and computational advantages not present in previous plug-in estimators. Specifically we will see that our methods admits an arbitrary number and type of confounders Z while estimating counterfactual probabilities, a stark difference to plug-in estimators, such as those presented in [CK20].

Our approach has two stages, training the neural network such that it learns the counterfactually ordered causal mechanisms that best fit the data, then interpreting it as a twin network on which standard inference is performed to estimate counterfactual distributions. Note that if one can generate counterfactuals, one can also generate interventions.

For clarity, we confine our explanations to the causal structure from Figure 2.8a where X, Y are categorical variables with $X \in \{1, \ldots, N\}$ and $Y \in \{1, \ldots, M\}$, and Z can be categorical or numerical. Note there can be many Z. Our method can be extended to multiple causes and a single output straightforwardly. To generalize this approach to an arbitrary causal structure, one applies our method to each parent-child structure recursively in the topological specified by the direction of the arrows in the causal structure.

Training deep twin networks: To determine the architecture of our neural network, we start with the causal structure of the SCM we wish to learn, and consider the graphical structure of its twin network representation. Our neural network architecture then exactly follows this graphical structure. This is graphically illustrated for the case of binary X, Y from Figure 2.8a with twin network in Figure 2.8c in Figure 7.2. In the case of binary X, Y, the neural network has two heads, one for the outcome under the factual treatment and the other for the outcome under the counterfactual treatment. Furthermore two shared—but independent of one another—base representations, one corresponding to a representation of the observed confounders, Z, and the other to the latent noise term on the outcome, U_Y, are employed. For multiple treatments we have N neural network heads, each corresponding to the categories of X. To interpret this as a twin network for given evidence X, Y, Z and desired intervention X^*, we marginalize out the heads indexed by the elements of $\{1, \ldots, N\}/X, X^*$. To train this neural network, we require two things: 1) a label for head Y^*, and 2) a way to learn the distribution of the latent noise term U_Y.

For 1), we must ask what the expected value of Y^* is, for fixed covariates Z, under a change in input X^*. This corresponds to $\mathbb{E}(Y^*|X^*, Z)$. Given the correspondence between twin networks and the original SCM outlined in Equation (7.1) from Section 7.2.1, this corresponds to $\mathbb{E}(Y|do(X), Z)$, which is the expected value of Y under an intervention on X for fixed Z. There are many approaches to estimating this quantity in the literature [SJS16, AWVDS17, JSS16, SBV19]. We follow [SLK18]due to their method's simplicity and empirical high performance. Any method that computes $\mathbb{E}(Y|do(X), Z)$ can be used, however. In addition to specifying the causal structure, the following standard assumptions are needed to estimate $\mathbb{E}(Y|do(X), Z)$ [SLK18]: 1) *Ignorability:* there are no unmeasured confounders; 2) *Overlap:* every unit has

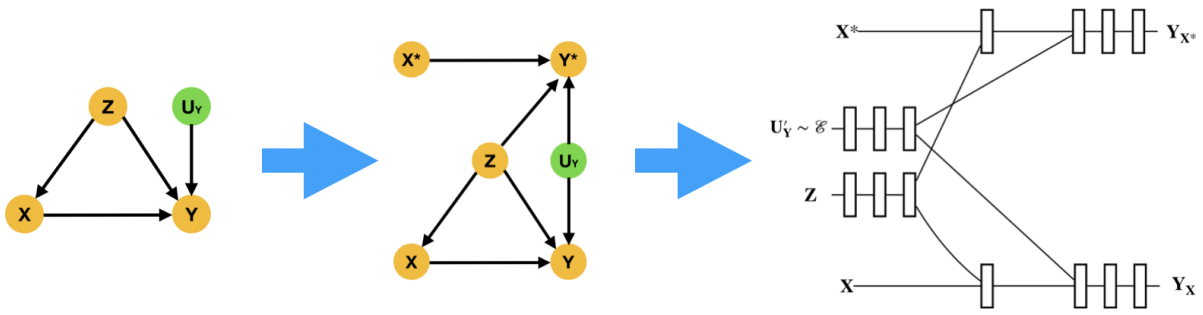


Figure 7.2: From DAG to twin network DAG to deep neural network (NN) architecture for binary X, Y. Rectangular blocks are NN blocks, like FCN layers or Lattices; forward intersections are concatenation of features.

non-zero probability of receiving all treatments given their observed covariates. Computing this expectation provides the labels for Y^*.

For 2), consider the following. Formally, the causal structure of Figure 2.8a has $Y = f(X, Z, U_Y)$ with $U_Y \sim q(U_Y)$ for some q. Without loss of generality [GKC+18], one can rewrite this as $Y = f(X, Z, g(U'_Y))$ with $U'_Y \sim \mathcal{E}$ and $U_Y = g(U'_Y)$, where $\mathcal{E}$ is some easy-to-sample-from distribution, such as a Gaussian or Uniform. Hence we have reduced learning $q(U_Y)$ to learning function g, whose input corresponds to samples from a Guassian or Uniform. Taken together, this provides a method to train our deep twin network. A summary is provided in Algorithm 2.

Algorithm 2 Training a deep twin network

Input: X: Treatment, Z: Confounders, X^*: Counterfactual Treatment; Y: Outcome; C: DAG of causal structure; I: loss imposing constraint on causal mechanisms
Output: F: trained deep twin network

1: Set F's architecture to match twin network representation of C, as in Figure 7.2
2: To obtain label for counterfactual head, estimate $\mathbb{E}(Y|do(X), Z)$, yielding training dataset $\mathcal{D} := \{X, X^*, Z; Y, Y^*\}$
3: **for** $x, x^*, z; y, y^* \in \mathcal{D}$ and $u_y \sim \mathcal{N}(0, 1)$ **do**
4: $\quad y', y'^* = F(x, x^*, u_y, z)$
5: $\quad$ Train F by minimizing $MSE(y, y') + MSE(y^*, y'^*) + I(\mathcal{D})$
6: **end for**

Enforcing constraints on the causal mechanisms: There are a few approaches to ensure that the function space learned by a neural network satisfies counterfactual ordering. Recall from Theorem 7.1 that such constraints correspond to limits on the type of input-output pairs consistent with the function. One approach is to specify a loss function penalising the network for outputs that violate the constraints, as done in [SAM97]. Alternatively, counterexample-guided learning [SFMB20] can be employed, to ensure the trained network does not produce any of these outputs when given the corresponding input. Lastly, as Theorem 7.3 equates

counterfactual ordering with monotonicity, a recent method uses "look-up tables" [GCP+16] to enforce monotonicity, and has been implemented in TensorFlow Lattice. Theorem 7.3 requires that the treatment and outcome categorical variables be themselves ordered. One method of ordering our treatment variables is based on their perceived preference, i.e. based on domain knowledge. In the absence of domain knowledge one could look at the Average Treatment Effect (ATE) =: $\frac{1}{N}\sum_i(y_1(i) - y_0(i))$, where y_1 represents the treated outcome while y_0 the control outcome, as well as the interventional probabilities $P(Y \mid do(X))$. We then determine an estimate of the relationship governing the treatment and the outcomes by observing the trend of the ATE or $P(Y \mid do(X))$ as we change the treatments. The treatments are then ordered such that their relationship towards the outcomes can be characterized as monotonic. If one does not have access to interventional probabilities, then one can estimate them from observational data [SLK18].

Estimating counterfactuals: We can now use the trained model to perform counterfactual inference. The reason our neural network architecture matches the Twin Network structure is that performing Bayesian inference on the neural network explicitly equates to performing counterfactual inference. For Figure 2.8a, there are two counterfactual queries one can ask: (1) $P(Y_{X=x'} = y' \mid X = x, Y = y, Z = z)$, (2) $P(Y_{X=x} = y, Y_{X=x'} = y' \mid Z = z)$, where any of x, y, z can be the empty set. Recall from Section 7.2.1 that in a twin network (1) corresponds to $P(Y^* = y' \mid X = x, Y = y, X^* = x')$, and (2) corresponds to $P(Y = y, Y^* = y' \mid X = x, X^* = x')$. Any method of Bayesian inference can then be employed to compute these probabilities, such as Importance or Rejection Sampling, or Variational methods. See Algorithm 3.

Algorithm 3 Deep twin network counterfactual inference

Input: X: Treatment, U_Y: Noise, Z: Confounders, X^*: Counterfactual Treatment; Y: Outcome Y^*: Counterfactual Outcome; F : Trained deep twin network; Q: desired counterfactual query (in this example, $Y_{X=x'} = y' \mid X = x, Y = y, Z = z)$)
Output: P(Q): Estimated distribution of Q.

Convert P(Q) to twin network distribution: $P(Y_{X=x'} = y' \mid X = x, Y = y, Z = z) \rightarrow P(Y^* = y' \mid X = x, Y = y, X^* = x')$
Compute $P(Y^* = y' \mid X = x, Y = y, X^* = x')$:
for $x, x', z \in \mathcal{D}_{test}$ **do**
    for $[u_y]_N \sim \mathcal{N}(0, 1), N \in \mathbb{N}$ **do**
        Sample $(\tilde{y}, \tilde{y}^* = F(x, x', u_y, z))$ such that $\tilde{y} = y$
        The frequency of these samples for which $\tilde{y}^* = y'$ yields P(Q)
    end for
end for

7.3.6 Probabilities of Causation: Definitions

The probabilities of causation are important counterfactual queries that quantify the degree to which one event was a necessary or sufficient cause of another. Recently these have been applied in medical diagnosis [RLJ20] to determine if a patient's symptoms would not have occurred had it not been for a specific disease. Here, the proposition binary variable W is true will is denoted $W = 1$, while its negation, $W = 0$, denotes the proposition W is false.

1. **Probability of necessity:**

$$P(Y_{X=0} = 0 \mid X = 1, Y = 1)$$

The probability of necessity is the probability event Y would not have occurred without event X occurring, given that X, Y did in fact occur.

2. **Probability of sufficiency:**

$$P(Y_{X=1} = 1 \mid X = 0, Y = 0)$$

The probability of sufficiency is the probability that in a situation where X, Y were absent, intervening to make X occur would have led to Y occurring.

3. **Probability of necessity & sufficiency:**

$$P(Y_{X=0} = 0, Y_{X=1} = 1 \mid Z)$$

The probability of necessity & sufficiency quantifies the sufficiency and necessity of event X to produce event Y in context Z. As discussed in section 2.9.2, joint counterfactual probabilities are well-defined.

7.4 Related Works

Despite the large body of work using machine learning to estimate interventional queries relatively little work has explored using machine learning to estimate counterfactual queries.

Machine learning to estimate counterfactual queries: Recent work from [PCG20] used normalising flows and variational inference to compute counterfacual queries using abduction-action-prediction. A limitation of this work is that identifiability constraints required for the counterfactual queries to be uniquely defined given the training data are not imposed. Work by Oberst and Sontag [OS19], expanded by [LJM$^+$21], used the Gumbel-Max trick to estimate counterfactuals, again using abduction-action-prediction. While this methodology satisfies generalisations of the monotonicity constraint, it does so because the Gumbel-Max trick has a limit on the type of conditional distributions it can generate—not because the authors imposed partial-identfiability constraints during the learning process. Hence the Gumbel-Max may not be suitable for the computation of counterfactual queries requiring different (partial-)identifiability constraints. Additional work by Cuellar and Kennedy, [CK20] devised a non-parametric method to compute the Probability of Necessity using an influence-function-based estimator. This estimator was derived under the assumption of monotonicity. A limitation of this approach is that a separate estimator must be derived and trained for each counterfactual query.

A large body of work addresses the issue of *partial* identifiability of counterfactuals from data. Historically, this line of work was initiated by [BP97], who explored bounds on the probabilities of causal queries of binary variables using linear programming. Recently, [ZB20] extended such linear programming derived upper and lower bounds beyond binary outcomes to the case of continuous outcomes. Additional work by [JZ21] bounded counterfactuals by mapping the SCM space onto a new one that is discrete and easier to infer upon. Finally, [IA94] proposed Local average treatment effects as means of identifications of interventional queries from observational data.

Machine learning to estimate interventional queries: Recently there has been much

interest in using machine learning to estimate interventional conditional distributions. These were aimed at learning conditional average treatment effects: $\mathbb{E}(Y_{X=1} \mid Z) - \mathbb{E}(Y_{X=0} \mid Z)$. Examples include PerfectMatch [SLK18], DragonNet [SBV19], PropensityDropout [AWVDS17], Treatment-agnostic representation networks (TARNET) [SJS16], Balancing Neural Networks [JSS16]. Each work utilised neural network architectures similar to the one depicted in Figure 7.2, with sight modifications based on the specific approach. Other machine learning approaches to estimating interventional queries made use of GANs, such as GANITE [YJVDS18] and CausalGAN [KSDV18], Gaussian Processes [WTJM20, AvdS17], Variational Autoencoders [LSM$^+$17], and representation learning [ZBS20, AZT$^+$21, YLL$^+$18].

While our architecture shares similarities with these, there is one main difference. By interpreting our architecture as a twin network in the sense of [BP94]—as discussed in Section 7.3.5—and explicitly including an input for the latent noise term U_Y, we can elevate our network from estimating interventional queries to fully counterfactual ones by performing Bayesian inference on it.

7.5 Experimentation

We now evaluate our *counterfactual ordering* principle and our *deep twin network* computational tool. We focus on four publicly available real-world datasets, the German Credit Dataset [DG17], the International Stroke Trial (IST) [SN11], the Kenyan Water task [CK20] and the Twin mortality dataset [LSM$^+$17]. We further use two synthetic and two semi-synthetic tasks to test our proposed methods. Full dataset description is in Section 7.5.1.

We break our research questions (RQ) into two distinct types: ones that assess our counterfactual ordering principle, and ones that assess the counterfactual estimation accuracy of deep twin networks. Specifically, to assess counterfactual ordering, we wish to determine if not enforcing it leads to domain knowledge conflicting counterfactuals in real datasets, relative to enforcing it. To estimate counterfactual estimation accuracy of deep twin networks satisfying counterfactual ordering, we test on synthetic and semi-sythentic datasets—where we by design

have access to the ground truth—as well as on a dataset involving twins, where we use features of one twin as a the counterfactual for the other. Finally, we also test on a real dataset involving binary treatment and outcome. We do this as Corollary 1 showed that in counterfactually ordered causal models with binary variables, certain counterfactual probabilities are uniquely identified from data. Hence for binary variables we can determine how accurate our deep twin network method is relative to these known identified expressions.

- **RQ1:** If counterfactual ordering isn't enforced, do counterfactuals conflict with domain knowledge?
- **RQ2:** By imposing counterfactual ordering, do generated counterfactuals comply with domain knowledge?
- **RQ3:** Can we accurately estimate counterfactual probabilities using deep twin networks?

7.5.1 Description of Datasets Used

Semi-Synthetic examples for RQ1 & RQ2

Here we describe in detail all the datasets used in this chapter. In the German Credit Dataset the treatment is a four-valued variable corresponding to current account status, and the outcome is loan risk. The International Stroke Trial database was a large, randomized trial of antithrombotic therapy after stroke onset. The treatment is a three-valued variable corresponding to heparin dosage, and the outcome is a three-valued variable corresponding to different levels of patient recovery. In both cases we explore semi-synthetic settings, where the treatment and confounders are derived from the original dataset but the outcome is defined in a synthetic fashion. Synthetic outcomes allow us to determine the ground truth counterfactuals and probabilities of causation (see 7.3.6 for definition). We also evaluate our algorithm with the real world outcome of the German Credit Dataset.

Below we define the semi-synthetic outcome for both datasets - the German Credit Dataset

and the International Stroke Trials one:

$$Y = \begin{cases} X + Z & \text{if } U_Y = 0 \\ 0, & \text{if } U_Y = 1 \\ X * Z, & \text{if } U_Y = 2 \\ 2, & \text{if } U_Y = 3 \\ 1, & \text{if } U_Y = 4 \\ \text{step}(X - 1), & \text{if } U_Y = 5 \\ 2 * \text{step}(X - 1), & \text{if } U_Y = 6 \end{cases} \tag{7.3}$$

where the treatment X and the confounders Z span the range $X, Z \in [0, 2]$. Step is the Heaviside step function. In our experimentation U_Y was drawn from a uniform distribution

Meanwhile for the IST dataset: Let X be the dosage of aspirin treatment and the counfouders be defined as SEX := biological sex of patient, AGE:= age of patient thresholded at 71 years, CONSC := level of consciousness the patient arrived in hospital with. The outcome $Y = 1/(1 + e^{-g})$ with g being given by:

$$\begin{aligned} g = X + \text{SEX} + 0.2 * (\text{CONSC} - 1) \\ + 0.5 * X * \text{SEX} * \text{AGE} + U_y \end{aligned} \tag{7.4}$$

Real World examples for RQ3

Moreover, the Kenyan Water task is to understand whether protecting water springs in Kenya by installing pipes and concrete containers reduced childhood diarrhea, given confounders. Firstly, monotonicity is a reasonable assumption here as protecting a spring is not expected to increase the bacterial concentration and hence increase the incidence of diarrhea. [CK20] reported a low value for Probability of Necessity here—suggesting that children who developed diarrhea after being exposed to a high concentration of bacteria in their drinking water would have contracted the disease regardless. However, as there is no ground truth here, further studies reproducing

this result with alternate methods are required to gain confidence in [CK20]'s result. We follow the same data processing as in [CK20], The Kenyan water dataset originates from [KLMP15] licensed under a non commercial use clause and with the requirement for secure storage, both conditions have been fulfilled by the authors. The data was preprocessed following [CK20].

For the Twin Mortality data, two versions were used. First databases provided by [LSM+17] were processed to remove NaNs. No further processing was administered. This constituted the completely real version of the Twin Mortality dataset. However, as both [LSM+17] and [YJVDS18] process the data to create a semi-synthetic task, in the spirit of proper comparison we used the data as processed and provided by [YJVDS18], with no additional processing.

Synthetic examples for RQ3

Finally we set out the two synthetic examples used in RQ3 to test our Deep Twin Network methodology. Let:

$$Y = \begin{cases} X & \text{if } U_Y = 0 \\ 0, & \text{if } U_Y = 1 \\ 1, & \text{if } U_Y = 2 \end{cases} \tag{7.5}$$

Hence

$$P(N) = \frac{P(U_Y = 0)}{P(U_Y = 2) + P(U_Y = 0)} \tag{7.6}$$

$$P(S) = \frac{P(U_Y = 0)}{P(U_Y = 1) + P(U_Y = 0)} \tag{7.7}$$

$$P(NS) = P(U_Y = 0) \tag{7.8}$$

In addition we consider a confounded example for RQ3

$$Y = \begin{cases} X \times Z, & \text{if } U_Y = 0 \\ 0, & \text{if } U_Y = 1 \\ 1, & \text{if } U_Y = 2 \end{cases} \tag{7.9}$$

ATE	0	1	2
0	0	0.3059	0.8914
1	-0.3059	0	0.5854
2	-0.8914	-0.5854	0

Table 7.1: Treatment: Semi-Synthetic Existing account status, Outcome: Synthethic. As the change from treatment 0 to 1&2 has a positive Average Treatment Effect, the relationship is increasing monotonic. Here we are investigating our ability to determine the direction of monotonicity. We observe that the ATE is positive as we increase the treatment we get a larger ATE, hence the monotonicity direction is correctly surmised

where

$$X := U_x \oplus Z$$

Hence

$$P(N) = \frac{P(U_Y = 0)P(Z = 1)}{P(U_Y = 2) + P(U_Y = 0)P(Z = 1)} \tag{7.10}$$

$$P(S) = \frac{P(U_Y = 0)P(Z = 1)}{P(U_Y = 1) + P(U_Y = 0)} \tag{7.11}$$

$$P(NS) = P(U_Y = 0)P(Z = 1) \tag{7.12}$$

7.5.2 Determining monotonicity direction

Before we attempt to answer the aforementioned Research Questions we investigate methods to determine the monotonicity direction at a given dataset. In Table 7.1 we provide the ATE of a semi-synthetic existing account status from the German Credit Dataset [DG17] with a fully synthetic outcome defined in Equation (7.3). We observe that for a control treatment 0 changing the treatment to $[1, 2]$ the ATE increases, indicating a monotonic increasing relationship. In addition, we could observe the interventional probabilities where as we increase the value of the treatment the probability of a higher outcome increases, we show this in Table 7.2. This reinforces our beliefs regarding the type of monotonicity.

Similarly, Table 7.3, Table 7.4 include the ATE and the interventional probabilities for a real world variant of the above dataset where the treatments are again the current account status of the individual but the outcome is their classification as good or bad risk. At this point, we

$P(Y\|do(X))$	**0.0**	**1.0**	**2.0**
0	0.6396	0.2056	0.1548
1	0.4635	0.2518	0.2847
2	0.1656	0.2620	0.5723

Table 7.2: Same data as Table 7.1. Here we look at the interventional probabilities. As Rows are treatments, and columns are outcomes

ATE	**0**	**1**	**2**	**3**
0	0	-0.0791	0.1029	0.2174
1	0.0791	0	0.1820	0.2965
2	-0.1029	-0.1820	0	0.1145
3	-0.2174	-0.2965	-0.1145	0

Table 7.3: Following the same investigation as Table 7.1. Treatment: Account status, Outcome: Risk Status. We get the same insights but there is a potential we could be exchanging Treatment 0 and 1, As we see later on this is not necessary and we hypothesise this violation in the monotonic ATE is due to noise of the real world data.

may call upon our domain knowledge and determine if the break in the monotonic trend is due to noisy observations or a different ordering of the treatments. As this is a real world dataset in which the outcome attribution is inherently noisy we observe an outlier behavior from treatment 1. Upon closer inspection we observe that treatment 1 corresponds to a negative balance in the individual's checking account while treatment 0 indicates no existing checking account. As such one could either switch the treatment ordering to obey the monotonicity, or in the case that this break in monotonicity is suspected to be due to noisy data one could enforce prior knowledge-based monotonicity. Here, we follow our prior knowledge and attribute the break of monotonicity to noise. Our reasoning is based on the fact that an individual without a prior credit account is a larger unknown for a financial institution.

$P(Y\|do(X))$	**0.0**	**1.0**
0	0.1919	0.8081
1	0.5450	0.4549
2	0.3336	0.6663
3	0.1265	0.8735

Table 7.4: $P(Y|do(X))$ of the same dataset, rows indicate treatments while columns outcomes

P(T',T)	0.0	1.0	2.0	3.0
0	0	0.0816 ± 0.1414	0.0860 ± 0.1191	0.0344 ± 0.0208
1	0.1439 ± 0.1396	0	0.1156 ± 0.0984	0.1297 ± 0.1306
2	0.1286 ± 0.0726	0.1290 ± 0.1347	0	0.0741 ± 0.0752
3	0.0680 ± 0.0457	0.1854 ± 0.1452	0.0974 ± 0.1140	0

Table 7.5: **Non-constrained model**. $P(T', T) = P(\text{Risk}_{\text{Account Status}=T'} = \text{good} \mid \text{Account Status} = T, \text{Risk} = \text{bad})$. Columns and rows are Treatments. We observe counter-intuitive probabilities as the lower triangular sub-matrix offers higher probabilities than the upper triangular one. That is, if we observe evidence where bad account status led to bad risk, the non-constrained model predicts an increase in net worth would have led to a *lower* chance of being deemed a good risk—even though all other factors are kept fixed. An un-intuitive result that conflicts with domain knowledge of the finance industry.

P(T',T)	0.0	1.0	2.0	3.0
0	0	0.3022 ± 0.0415	0.3977 ± 0.0382	0.4040 ± 0.0381
1	0.1079 ± 0.0322	0	0.3891 ± 0.0988	0.4118 ± 0.0545
2	0.0670 ± 0.0156	0.2816 ± 0.0653	0	0.4470 ± 0.0751
3	0.1383 ± 0.0442	0.2953 ± 0.0344	0.3522 ± 0.0577	0

Table 7.6: **Counterfactual Ordering**. $P(T', T) = P(\text{Risk}_{\text{Account Status}=T'} = \text{good} \mid \text{Account Status} = T, \text{Risk} = \text{bad})$. Columns and rows are Treatments. We observe intuitive results as the lower triangular sub-matrix offers lower probabilities than the upper triangular one. That is, when we observe evidence in which bad account status led to bad risk, the counterfactually ordered model predicts an increase in net worth would have led to a higher chance of being deemed a good risk—an intuitive result that complies with domain knowledge in the finance industry.

7.5.3 Answering RQ1 & RQ2

We now investigate the German Credit real-world dataset and the International Stroke Trial dataset in an effort to answer RQ1 and RQ2. We train a deep twin network on German Credit data using Algorithm 2.

In Table 7.6 and Table 7.5 we estimate the counterfactual probability $P(\text{Risk}_{\text{Account Status}=T'} = \text{good} \mid \text{Account Status} = T, \text{Risk} = \text{bad})$ for a model satisfying counterfactual ordering and an unconstrained model respectively. That is, we ask what the probability that our loan risk would be good if we improved our account status, given that our account status is currently bad and we were just deemed a poor risk of a loan. We note that the unconstrained model offers us non-intuitive probabilities that, when put in context, do not make sense in the real

P	0.0	1.0	2.0	3.0
0	0	0.4773 ± 0.0182	0.4243 ± 0.0272	0.3894 ± 0.0147
1	0.6049 ± 0.0386	0	0.5791 ± 0.0340	0.5616 ± 0.0183
2	0.6081 ± 0.0388	0.6025 ± 0.0806	0	0.5221 ± 0.0130
3	0.6265 ± 0.0297	0.6580 ± 0.0431	0.5188 ± 0.0272	0

Table 7.7: Switched counterfactual ordering – Probability of counterfactual $P(T, T') = P(Y_{X=T'} = 1 \mid X = T, Y = 0)$ – columns and rows are Treatments – We observe counter-intuitive probabilities of necessity as the lower triangular sub-matrix has higher probabilities than the upper triangular

P	1.0	2.0	
0	0	0.1260 ± 0.0070	0.0000 ± 0.0000
1	0.0000 ± 0.0000	0	0.2262 ± 0.0429
2	0.0000 ± 0.0000	0.0000 ± 0.0000	0

Table 7.8: $P = P(Y_{X=1} = Column | Y_{X=0} = Row)$. In this semi-synthetic example we expected a good behaving model that provides the lower triangular part of the table to be 0.

world. We observe that when we condition on evidence in which bad account status led to bad risk, the unconstrained model predicts that increasing an individual's net worth would have resulted in a lower probability of being deemed a good risk than *decreasing* their net worth. This result defies common sense, answering RQ1. On the other hand, when we observe bad account status led to bad risk, the counterfactually ordered model predicts that an increase in an individuals net worth would have led to a higher chance of them being deemed a good risk than a decrease in their worth in this context. This result fits with our understanding of the financial industry, answering RQ2.

In Table 7.7 we show a special ablation case where we on purpose changed the direction of monotonicity leading to the generation of counterfactuals that clash with our understanding of the financial industry - further supporting our argumentation.

Moreover, Tables 7.8,7.9 show counterfactual probabilities from our method applied to the semi-synthetic account status treatment and synthetic outcome. We note that our model slightly violates the monotonicity constraints by providing non zero probabilities to two cases where they should be 0. However, both of these are within our acceptable experimental error, with one being less than 1%, the other being just over 1%

P	0.0	1.0	2.0
0	0.0000	0.1190	0.0079
1	0.0000	0.0000	0.2037
2	0.0000	0.0000	0.0000

Table 7.9: $P = P(Y_{X=1} = Column|Y_{X=0} = Row)$.

P	1.0	2.0	
0	0	0.0482 ± 0.0006	0.1840 ± 0.0001
1	0.0011 ± 0.0019	0	0.1130 ± 0.0069
2	0.0000 ± 0.0000	0.0135 ± 0.0058	0

Table 7.10: $P = P(Y_{X=1} = Column|Y_{X=0} = Row)$.

Finally regarding the IST dataset, Tables 7.10,7.11 show the counterfactual probabilities for the semi-synthetic heparin treatment and synthetic outcome.

7.5.4 Answering RQ3

As a reminder we reiterate the third research question, **RQ3:** Can we accurately estimate counterfactual probabilities using deep twin networks?

Synthetic data We first evaluate whether we can accurately estimate the probabilities of causation on synthetic data. We test on data generated by an unconfounded as well as a confounded synthetic causal model, whose functional forms are outlined in the Section 7.5.1. Following the algorithms 2, 3 we train a deep twin network on data from each case and enforce monotonicity.

We test on both unconfounded and confounded causal models, with causal structure from Figure 2.8a and Figure 2.8e respectively. The data generation functions are defined in Equation (7.9) and Equation (7.5) The functions remain monotonic in X. Given these, we construct synthetic datasets of 200000 points split into training and testing under an $80 - 20$ split. The samples U_y were drawn from either a uniform or a Gaussian distribution, depending on the experiment. Confounders Z were taken from a uniform distribution. We opt for a high number of samples such that we do not bias our analysis due to small sample sizes. In the real world

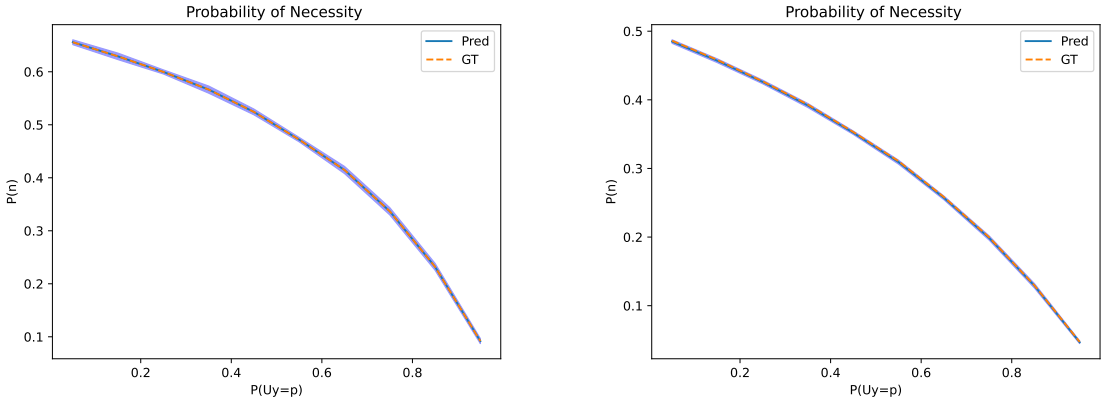
P	0.0	1.0	2.0
0	0.0000	0.0266	0.0917
1	0.0000	0.0000	0.0911
2	0.0000	0.0000	0.0000

Table 7.11: $P = P(Y_{X=1} = Column | Y_{X=0} = Row)$.

Method	U_y	P(N)	P(S)	P(N&S)
Synth Ground Truth	Uniform	0.5	0.5	0.33333
Synth Twin Net	Uniform	0.50214 ± 0.00387	0.50046 ± 0.00631	0.33449 ± 0.00401
Synth w/ Conf Ground Truth	Gaussian	0.54706	0.35512	0.27443
Synth w/ Conf Twin Net	Gaussian	0.54563 ± 0.00276	0.35177 ± 0.00144	0.27207 ± 0.00125

Table 7.12: Results of Synthetic experiments. P(N): Prob. of Necessity; P(S): Prob. of Sufficiency; P(N&S): Prob. of Necessity and Sufficiency. Our model achieves highly accurate estimations of the probabilities of causation on synthetic data.

experiments the dataset sizes are smaller. Results for a trained twin network are in Table 7.12. We accurately estimate all Probabilities of Causation in both unconfounded and confounded cases when ground truth and candidate distributions are the same. In Figure 7.3 we also show performance of (a) unconfounded and (b) confounded cases as ground truth distribution of U_Y in synthetic generating functions changes, but candidate training distributions remain fixed—showing robust estimation.



(a) (b)

Figure 7.3: Predicted & ground truth Probability of Necessity as distribution of U_Y varies in synthetic generating functions, but training distributions do not. Plots show robust estimation. (a) unconfounded, (b) confounded. Errors bars in both

Method	P(N)	P(S)	P(N&S)	AUC-ROC / F1
KW Median Child *Cuellar et al. 2020*	0.12 ± 0.01	-	-	-
KW TN Median Child	0.13598 ± 0.049	0.09811 ± 0.031	0.31778 ± 0.012	-
KW TN Test Set	0.06273 ± 0.020	0.03914 ± 0.016	0.08521 ± 0.034	-
Twin Mortality Ground Truth	0.33372	0.01011	0.01353	0.83/- *Louizos et al. 2017*
TM TN Test Set	0.12241 ± 0.019	0.01401 ± 0.003	0.01174 ± 0.002	0.86/0.83

Table 7.13: Results of Kenyan Water (KW) & Twins Mortality (TM) with Twin Network (TN), P(N): Prob. of Necessity; P(S): Prob. of Sufficiency; P(N&S): Prob. of Necessity & Sufficiency. In KW we agree & improve on [CK20]. In TM we overestimate P(N), but report accurate P(S) & P(N&S), & better AUC than [LSM$^+$17].

	Credit Dataset		IST - Aspirin		IST - Heparin	
F1 Scores	No Constrains Linear Layers	Counterfactual Ordering	No Constrains Linear Layers	Counterfactual Ordering	No Constrains Linear Layers	Counterfactual Ordering
Factual	0.4929	0.8637	0.6113	0.6417	0.3497	0.9758
Counterfactual	0.4698	0.9795	0.7152	0.9501	0.4103	0.9851

Table 7.14: F1 score of counterfactual predictions for semi-synthetic German Credit Dataset with Treatment: Existing account status, Outcome: Synthetic; & International Stroke Trial (IST) Dataset with Treatment: Aspirin, Outcome: Synthetic; Treatment: Heparin, Outcome: Synthetic. See Section 7.5.1 for dataset description.

Semi-synthetic data In Table 7.14 we show the results of our semi-synthetic experiments, for both the German Credit Datasest as well as the International Stroke Trial. Here, as the outcomes are synthetic, the ground truth is known a priori, hence we are able to calculate the associated F1 scores for each of the models. The counterfactually ordered models are more accurate at predicting both factual and counterfactual outcomes answering RQ3. Moreover, not enforcing counterfactual ordering leads to reduced performance in counterfactual estimation.

Real-world data We show now, and discuss performance on the Twin Mortality data of [LSM$^+$17] and the Kenyan Water task from [CK20].

In the case of Kenyan Water dataset, treatment and outcome are binary, so we can compare the deep twin networks estimated counterfactual distributions to the uniquely identified counterfactual distribution via Corollary 1. The Kenyan Water task is to understand whether protecting water springs in Kenya by installing pipes and concrete containers reduced childhood diarrhea. First, monotonicity is reasonable here as protecting a spring is not expected to increase the bacterial concentration and hence increase the diarrhea incidence. [CK20] reported a low value for Probability of Necessity here—suggesting that children who developed diarrhea after being

exposed to a high concentration of bacteria in their drinking water would have contracted the disease regardless. However, as there is no ground truth, further studies reproducing this result with alternate methods are required to gain confidence in [CK20]'s result. We follow the same data processing as in [CK20]. Our findings are in Table 7.13 and agree with [CK20] on Probability of Necessity. Moreover, unlike [CK20], we can also compute Probability of Sufficiency and Probability of Necessity and Sufficiency. We can thus offer a more comprehensive understanding of the role protecting water springs plays in childhood disease. Our results show that exposure to water-based bacteria is not a necessary condition to exhibit diarrhea and it is neither a sufficient, nor a necessary-and-sufficient condition. This provides further evidence that protecting water springs has little effect on the development of diarrhea in children in these populations, indicating the source of the disease is not related to water.

In the Twin mortality dataset the goal is to understand the effect being born the heavier of the twins has on mortality one year after birth, given confounders regarding the health of the mother and background of the parents. Previous work addressed this with intervention queries. We use counterfactual queries—specifically the probabilities of causation. We follow [LSM$^+$17, YJVDS18]'s preprocessing. As in [LSM$^+$17], we treat each twin as the counterfactual of their sibling—providing a ground truth reported in 7.13. Again, monotonicity is justified here as we do not expect increasing birth weight to lead to reduced mortality.
First, given birth weight and mortality evidence provided by one twin, we aim to estimate the expected counterfactual outcome had their weight been different. That is, compute

$$\mathbb{E}(\text{Mortality}_{\text{Weight}}|\text{Mortality}^*, \text{Weight}^*, Z),$$

where Z are observed confounders. We achieve a counterfactual AUC-ROC of 86% and F1 score of 83%. [LSM$^+$17] addressed this same question using only used interventional queries. That is, they computed $\mathbb{E}(\text{Mortality}_{\text{Weight}}|Z)$ and only achieve AUC 83%. We thus outperform [LSM$^+$17]'s AUC by 3%. Full results can be found in Table 7.13. Here, by explicitly conditioning on and using the fact that the observed twins had birth weight and mortality, we are able to update our knowledge about the latent noise term of the other twin. Our improved AUC score

showed using this allowed more accurate estimation of the "hidden" twins outcome. This cleanly illustrates the difference between interventions and counterfactuals. To give a comparison to prior work, we computed the average treatment effect from our model, yielding $-2.34\% \pm 0.019$ which matches [LSM+17]—showing our model accurately estimates interventions as well as counterfactuals.

Table 7.13 reports our estimation of the Probabilities of Causation. Note that no previous work has computed these counterfactual distributions. Despite accurate estimation of the Probability of Sufficiency, and Necessity & Sufficiency, our model underestimates the Probability of Necessity. This can be explained by a large data imbalance regarding the mortality outcome—affecting the Probability of Necessity the most as mortality is the evidence conditioned here. Nevertheless, we correctly reproduce the relative sizes of the Probabilties of Causation, with Probability of Necessity an order of magnitude larger than the others.

7.6 Summary

This chapter introduced *counterfactual ordering*, a principle that posits desirable properties causal mechanisms should posses to rule out "non-intuitive" counterfactuals. We proved it is equivalent to causal mechanisms being monotonic. To learn such causal mechanisms, and perform counterfactual inference with them, we introduced *deep twin networks*.

Throughout our experimentation we asked a series of counterfactual questions whose answers we found to be in accordance with the conclusions other pieces of literature have drawn. It is important to note though that all our results and methods assume the knowledge of a causal diagram. Determining such a diagram is not within the scope of causal inference but of causal discovery. Such tasks are interesting and challenging. Moreover, as we are unable to determine experimentally the validity of our counterfactual predictions we would rely on our mathematical analysis and other methods like sensitivity analysis and refutation methods to determine the performance of methods.

Chapter 8

Counterfactual Inference in Medical Images

As we explored topics in causally enabled ML and its application to medical data, we developed a novel methodology - the deep twin networks. In this chapter we look into applying our method to medical images. We further tackle a real world limitation of lack of total identifiability. As medical images and phenomena are high dimensional and complex the standard identifiability constraints cannot be applied. The need, thus, arises to make expert knowledge judgments to determine the validity of the learned models. This work has been done in collaboration with Hadrien Reynaud that is the first author of the associated publication [RVH+21], featured in MICCAI 2022. I contributed with the idea of applying the deep twin networks from Chapter 7, the theory of twin networks and a baseline implementation of them in Tensorflow 2. H.R. then re-implemented them in Pytorch and engineered the pipeline required to train them with images. I also provided continual support by guiding H.R. to potential solution for the latter part of the project where access to true counterfactuals was not possible.

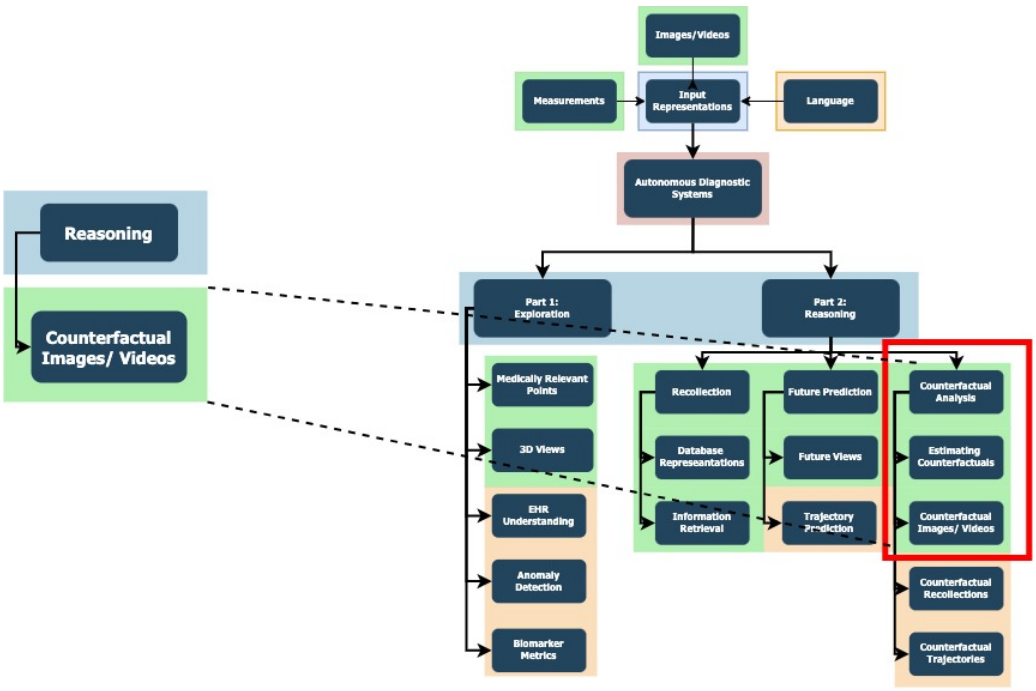


Figure 8.1: In this chapter we look into how to perform counterfactual image and video generation

8.1 Introduction

How would this patient's scans look like if they had a different LVEF? How would this US view appear if I turned the probe by 5 degrees? These are important causality related questions that physicians and operators ask explicitly or implicitly during the course of an examination in order to reason about the possible pathologies of the patient. While in the second case the interventional query of turning the probe is easy to resolve – by performing the action– queries like the first case, where we ask a counterfactual question, cannot be answered that easily. Indeed, these fall under the third and highest rung of Pearl's [Pea09] hierarchy of causation.

Counterfactual queries probe into alternative scenarios that might have occurred had our actions been different. For the first question of this paper, we ask ourselves how the patients scans would look like if they had a different LVEF. Here, the treatment would be the different ejection fraction, and the outcome, the different set of scans. Note that this is a query that is *counter-to* our observed knowledge that the patients scans exhibited a specific LVEF. As such, standard Bayesian Inference that conditions on the observed data without any further considerations is not able to answer this type of questions.

8.1.1 Related works

Generating synthetic US images can be performed with physics-based simulators [SHN08, CYLW13, MMG14, GCC+09, BBRH12] and other techniques, like registration-based methods [LCKD+05]. However, these methods are usually very computationally expensive and do not generate fully realistic images. With the shift into deep learning, GAN-based techniques have emerged. They can be based on simulated US priors or other imaging modalities (MRI, CT) [CFS20, TFY20, TGS+21, AASK+20, ASAL+20, TZPG21] to condition the anatomy of the generated US images. Recently, many works explore machine learning as a tool to estimate interventional conditional distributions [YJVDS18, KSDV18, LSM+17, AZT+21]. However, fewer works focus on the counterfactual query estimation. [PCG20, OS19] explore the Abduction-Action-Prediction paradigm and use deep neural networks for the abduction step, which is computationally very expensive. [CK20] derive a parametric mathematical model for the estimation of one of the probabilities of causation, while [VKGL21] use [BP94] to develop deep twin networks. The computer vision field also has a lot of interest for conditional generation problems. [WBBD20, TLYK18b] does conditional video generation from a video and a discrete class. [DWX+20] uses an image and a continuous value as input to produce a new image. [SG21, KSDV18] introduce causality in their generation process, to produce images from classes. More recently [CXL+] use diffusion models and the Abduction-Action-Prediction paradigm to generate synthetic aged cardiac images.

8.1.2 Contributions

In this chapter (1) We extend the causal inference methodology known as Deep Twin Networks Chapter 7 and[VKGL21] into a novel generative modelling method (D'ARTAGNAN [1]) able to handle counterfactual queries. (2) We apply our framework on the synthetic MorphoMNIST [CTK+19] and real-world EchoNet-Dynamic [OHG+20] datasets, exhibiting that our method can perform well in both fully controlled environments and on real medical cases. To the best of our knowledge, this is an entirely novel approach and task, and thus the first

[1] D'Artagnan is the fourth Musketeer from the French tale "The three Musketeers".

time such an approach is explored for medical image analysis and computer vision. Our work differentiates itself from all other generative methods, as it combines video generation with continuous conditional input, framed in a new causal framework that supports counterfactual queries to produce counterfactual videos with a semi-supervised approach allowing most standard labelled datasets to be used.

8.2 Method

Deep Twin Networks The methodology we propose is based on Deep Twin Networks. The training procedure and parametrization are borrowed from [VKGL21], who sets the foundation for such a causal framework. Deep Twin Networks use two branches, the factual and the counterfactual branch. We note our factual and counterfactual treatments (inputs unique to each branch) X and X^*, while the confounder (input shared between both branches) is noted Z. We note the factual and counterfactual outcomes $\hat{Y}, \hat{Y}^*$ while the noise, injected midway through the network, and shared by both branches, is noted U_Y. This information flow sets Z as the data we want to query with X and X^*, to produce the outcomes $\hat{Y}$ and $\hat{Y}^*$. These variables will be detailed on a case-specific basis in section 8.3. See Fig.8.2 for a visual representation of the information flow.

Synthetic data Full control over the data, as with synthetic data, enables the generation of both ground truth outcomes Y and Y^* along with their corresponding inputs X, X^* and Z. This makes training the Deep Twin Network trivial in a fully supervised fashion, we demonstrate in our first experiment.

Real-world data As we theoretically need paired data with precise labelling of their differences, applying our approach to the real-world might seem unfeasible as very few datasets are arranged as such. To overcome this limitation and enable our framework to support most standard labeled imaging dataset, we establish the following list of requirements that our model must possess to generate counterfactual videos for the medical domain: (1) produce a factual and counterfactual output that will share general visual features, such as style and anatomy, (2)

produce accurate factual and counterfactual videos with respect to the intervened upon variable, and (3) the counterfactual videos must be visually indistinguishable from real ones that possess the intervened upon features. In the following, we use the Echonet-Dynamic [OHG+20] dataset to illustrate the method, see Section 8.3 for details.

To resolve feature (1) we share the weights of the branches in the network, such that we virtually train a single branch on two tasks in parallel. To do so, we set the input video as our confounder Z and the labeled LVEF as the factual treatment X. We train the network to match the factual outcome $\hat{Y}$ with the input video. By doing so, the model learns to retain the style and anatomical structure of the echocardiogram in the confounder. This is presented as *Loss 1* in Figure 8.2 and mathematically defined as an L1 loss. For feature (2) we pre-train an expert network to regress the treatment values. The expert network takes an US video as input and outputs the estimated LVEF. The expert network weights are frozen when training the Twin model. It is used to train the counterfactual branch, for which we have no labeled data, thus preventing supervised training. The difference between the expert network prediction and the randomly sampled counterfactual treatment is denoted as *Loss 2* in Figure 8.2, also parameterized as an L1 loss. Finally, feature (3) calls for the well-known GAN framework, where we train a neural network to discriminate between real and fake images or videos, while training the Twin Network. This constitutes the adversarial *Loss 3* in Figure 8.2 that ensures that the counterfactual branch produces realistic-looking videos. With those 3 losses, we can train the generator (i.e. factual and counterfactual branches) to produce pairs of visually accurate and anatomically matching videos that respect their factual and counterfactual LVEFs treatment. In a scenario that identifiability criteria could be derived we would not be in need of the auxiliary losses, *Loss 3* and *Loss 1*. However general identifiability, and in particular in high dimensional cases, such guarantees are impossible to be given, [TP00]. Moreover, as neural networks are infamously good at deriving short cuts and not learning the features we theorize, we require the introduction of auxiliary losses.

To learn the **noise distribution** U_Y, we follow [GKC+17, VKGL21] and without loss of generality we can write $Y = f(X, Z, g(U'_Y))$ with $U'_Y \sim \mathcal{E}$ and $U_Y = g(U'_Y)$, where $\mathcal{E}$ is some easy-to-sample-from distribution, such as a Gaussian or Uniform. Effectively, we cast the prob-

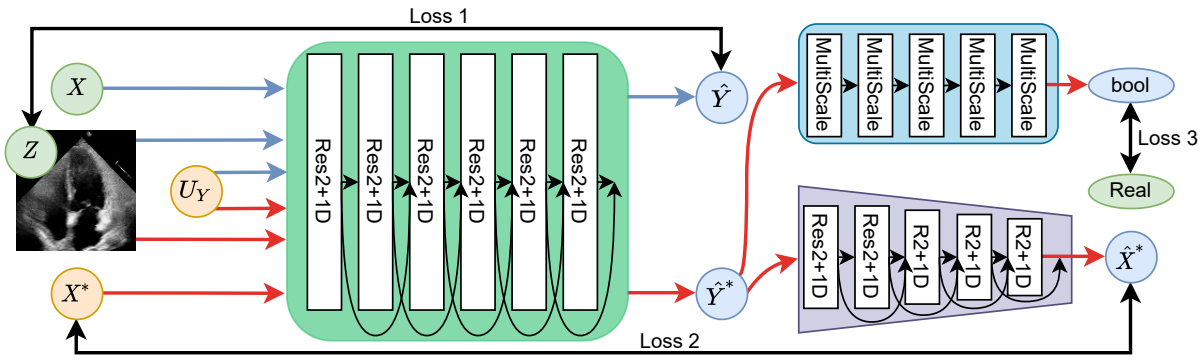


Figure 8.2: The D'ARTAGNAN framework. The green variables are known, the orange are sampled from distributions and the blue are generated by deep neural networks. Factual path is in blue, counterfactual in red.

lem of determining U_Y to learning the appropriate transformation from $\mathcal{E}$ to U_Y. For ease of understanding, we will be using U_Y henceforth to signify our approximation $g(U'_Y)$ of the true unobserved parameter U_Y. In addition to specifying the causal structure, the following standard assumptions are needed to correctly estimate $\mathbb{E}(Y|do(X), Z)$ [SLK18]: (1) *Ignorability:* there are no unmeasured confounders; (2) *Overlap:* every unit has non-zero probability of receiving all treatments given their observed covariates.

8.3 Experimentation

Datasets To evaluate D'ARTAGNAN, we use two publicly available datasets, the synthetic MorphoMNIST [CTK+19] and the clinical Echonet-Dynamic [OHG+20] dataset.

MorphoMNIST is a set of tools that enable fine-grained perturbations of the MNIST digits through four morphological functions, as wells as five measurement of the digits. To train our model, we need five elements: an original image, a (counter-)factual treatment $X(X^*)$ and a corresponding (counter-)factual label $Y(Y^*)$. To generate this data, we take 60.000 MNIST images I_i and sample 40 perturbation vectors $p_{i,j}$ for the five possible perturbations, including identity, thus generating 2.4 million images $I_{p_{i,j}}$. The perturbation vectors also encode the relative positions of the perturbations, when applicable. We *measure* the original images to produce vectors m_i and one-hot encode the labels into vectors l_i. We decided to perform

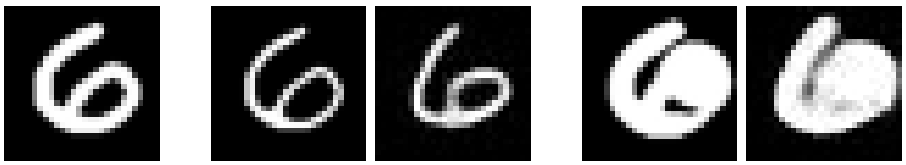


Figure 8.3: Left to right: original image, GT factual image, predicted factual image, GT counterfactual image, predicted counterfactual image. Factual perturbation is Thinning, Counterfactual perturbation are Thickening and Swelling.

the causal reasoning over a latent space, rather than image space. To do so, we train a Vector Quantized-Variational AutoEncoder (VQ-VAE) [OVK17] to project the MorphoMNIST images to a latent space $\mathcal{H} \in \mathbb{R}^{(q \times h \times w)}$ and reconstruct them. We opt for a VQ-VAE as it has been shown in literature to be quite powerful in terms of produced image quality, while maintaining a smaller memory and computational footprint than normalizing flows and diffusion models. Once trained, the VQ-VAE weights are frozen, and we encode all the ground-truth perturbed images $I_{p_{i,j}}$ into a latent embedding $\mathcal{H}_{i,j}$. Afterwards, the VQ-VAE is used to reconstruct the generated latent embeddings $\hat{Y}, \hat{Y}^*$ for qualitative evaluation purposes.

The clinical dataset Echonet-Dynamic [OHG+20] consists of 10,030 4-chamber echocardiography videos with 112×112 pixels resolution and various length, frame rates, image quality and cardiac conditions. Each video contains a single pair of consecutively labelled ES and ED frames. Each video also comes with a manually measured LVEF. For our use case, all videos are greyscaled and resampled to 64 frames, with a frame rate of 32 images per second. All videos shorter than two seconds are discarded, and we make sure to keep the labelled frames. For the resulting 9724 videos dataset, the original split is kept, with 7227 training, 1264 validation and 1233 testing videos.

MorphoMNIST For our synthetic experiment, we define a deep twin network as in [VKGL21] and follow the process we described in the Methods (Section 8.2). Regarding data organization, we use the elements defined in the section above. We set our confounder $Z_i = [l_i, m_i]$ to contain the one-hot encoded labels as well as the measurement of the original image. We sample two perturbations vectors $p_{i,m}$ and $p_{i,n}$ and their corresponding latent embeddings $\mathcal{H}_{i,m}$ and $\mathcal{H}_{i,n}$, where $n, m \in [\![0, 40[\![$, $n \neq m$. We set our input treatments as the perturbations vectors ($X = p_{i,m}$, $X^* = p_{i,n}$) and our ground-truth outcomes as the corresponding latent embeddings

Table 8.1: Metrics for MorphoMNIST (a) and EchoNet-Dynamic (b) experiments.

(a) MSE and SSIM scores. †No ordering is performed as there is no noise involved.

Metric	Factual	Counterfactual
MSE(Y, $\hat{Y}$)	2.3030	2.4232
SSIM(I_{gt}, I_{rec})†	0.9308	0.9308
SSIM(I_{rec}, I_{pred})	0.6759	0.6759
SSIM(I_{gt}, I_{pred})	0.6707	0.6705

(b) D'ARTAGNAN LVEF and reconstruction metrics.

Metric	Factual	Counterf.
R2	0.87	0.51
MAE	2.79	15.7
RMSE	4.45	18.4
SSIM	0.82	0.79

of the perturbed images ($Y = \mathcal{H}_{i,m}$, $Y^* = \mathcal{H}_{i,n}$). We sample $U_Y \sim [\mathcal{N}(0, 0.25) \bmod 1 + 1]$ and disturb the output of the branches of the neural networks that combine X and X^* with Z by multiplying the outputs of both with the same U_Y. With this setup, we generate factual and counterfactual perturbed MNIST embeddings from a latent description.

We assess the quality of the results by three means: (1) *Embeddings' MSE:* We sample a quintuplet (Z, X, X^*, Y, Y^*) and 1000 U_Y. The MSE between all $\hat{Y}_i$ and Y are computed and used to order the pairs ($\hat{Y}_i$, $\hat{Y}_i^*$) in ascending order. We keep the sample with the lowest MSE as our factual estimate and compute the MSE between $\hat{Y}_0^*$ and Y^* to get our counterfactual MSE score. (2) *SSIM:* We use the Structural SIMilarity metric between the perturbed images $I_{gt} = I_{p_{i,j}}$, the images reconstructed by the VQVAE I_{rec} and the images reconstructed from the latent embedding produced by the twin network I_{pred} to get a quantitative score over the images. (3) *Images:* We sample some images to qualitatively assess best and worst cases scenarios for this framework. We show the quantitative results in Table 8.1(a), and qualitative results in Figure 8.3.

Echonet Dynamic As stated in Section 8.2, our methodology requires an **Expert Model** to predict the LVEF of any US video. To do so, we re-implement the ResNet 2+1D network [TWT+18] as it was shown to be the best option for LVEF regression in [OHG+20]. We opt not to use transformers as they do not supersede convolutions for managing the temporal dimension, as shown in [RVH+21]. We purposefully keep this model as small as possible in order to minimize its memory footprint, as it will be operating together with the generator and the frame discriminator. The expert network is trained first, and frozen while we train the rest of D'ARTAGNAN.

D'ARTAGNAN We implement the generator as described in Section 8.2. We define a single deep network to represent both the factual and counterfactual paths. By doing so, we can meet the objectives listed in Section 8.2. The branch is implemented as a modified ResNet 2+1D [TWT+18] to generate videos. It takes two inputs: a continuous value and a video, where the video determines the size of the output. For additional details, please refer to fig. 8.2 and the code.

Discriminator We build a custom discriminator architecture using five "residual multiscale convolutional blocks", with kernel sizes 3, 5, 7 and appropriate padding at each step, followed by a max-pooling layer. Using multiscale blocks enables the discriminator to look both at local and global features. This is extremely important in US images because of the noise in the data, that needs to be both ignored, for anatomical identification, and accounted for to ensure that counterfactual US images look real. We test this discriminator both as a frame-based discriminator and a video-based discriminator, by changing the 2D layers to 3D layers where appropriate. We note that, given our architecture, the 3D version of the model requires slightly less memory but doubles the processing power compared to the 2D model.

Training the framework At each training step, we sample an US video (V) and its LVEF (ψ). We set our factual treatment, $X = \psi$ and our counterfactual treatment $X^* \sim \mathcal{U}(0, \psi - 0.1) \cup \mathcal{U}(\psi + 0.1, 1)$. The confounder Z and the factual ground truth Y are set to the sampled video such that $Z = Y = V$. We compute an L1 reconstruction loss (loss 1) between $\hat{Y}$ and $Y = V$. As we do not have ground truth for the counterfactual branch, we use the frame discriminator and the expert model to train it. Both models take as input the counterfactual prediction $\hat{Y}^*$. The expert model predicts an LVEF $\hat{\psi}$ that is trained to match the counterfactual input X^* with an L1 loss (loss 2). The discriminator is trained as a GAN discriminator, with $\hat{Y}^*$ as fake samples and V as real samples. It trains the generator with L1 loss (loss 3). The discriminator and expert model losses are offset respectively by three and five epochs, leaving time for the generator to learn to reconstruct V, thus maintaining the anatomy and style of the confounder. Our experiments show that doing so increases the speed at which the network is capable of outputting realistic-looking videos, thus speeding up the training of the discriminator that sees accurate fake videos sooner. Once the generator and discriminator are capable of generating

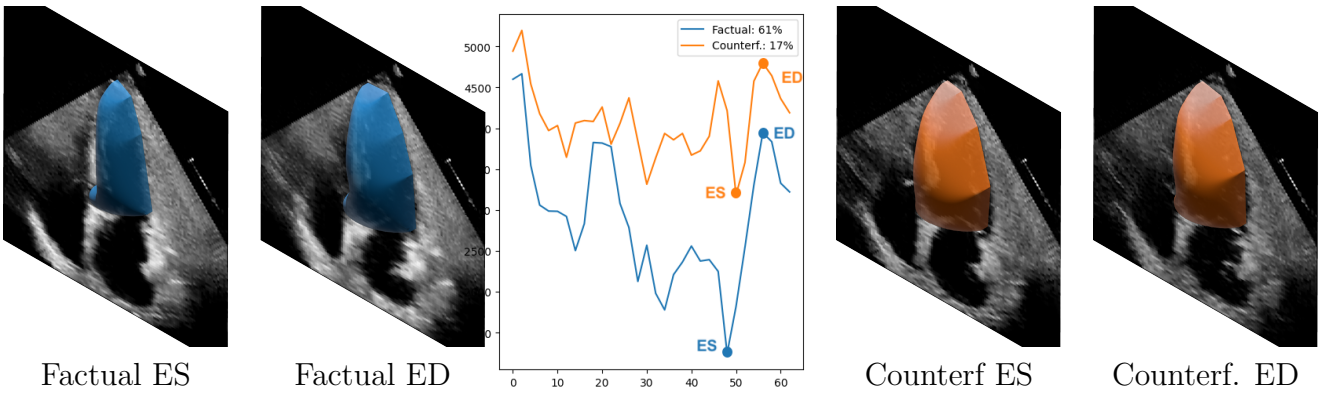


Figure 8.4: Qualitative results for D'ARTAGNAN over the same confounder and noise. Left: factual ES and ED frames. Middle: left ventricle area over time, obtained with a segmentation network as in [OHG+20]. Dots represent where the corresponding frames were sampled. Right: counterfactual ES and ED frames. Anatomy is preserved across videos, while the LVEF fraction is different.

and discriminating realistic-looking videos, the expert network loss is activated and forces the generator to take into account the counterfactual treatment, while the factual treatment is enforced by the reconstruction loss. The losses are also scaled, such that the discriminator loss has a relative weight of 3 compared to the reconstruction and expert loss.

Metrics To match our three objectives, we evaluate this new task on 1) the accuracy of the anatomy and 2) the accuracy of the regressed LVEF. We obtain the best possible video by sampling 100 U_Y and keep the $\hat{Y}^*$ with $\hat{\phi}^*$ closest to X^*. We then evaluate our metrics over those "best" videos. The anatomical accuracy is measured with SSIM and the LVEF precision is measured using R2, MAE and RMSE scores. Results are shown in Tables 8.1b, 8.2 and 8.3. We especially note, Table 8.3 where we compare our method with other non causal methods with respect to the output video quality. We observe that a causal method produces higher quality, in terms of SSIM, results. As such we expect that the generated LVEF would also be more accurate in the causal methodology. Further support for the aforementioned statement can be seen in the per frame MAE and RMSE seen in Table 8.2 that is equivalent or surpasses non causal methods. Such a result does agree with our intuition that a causal model is more robust and has learned the true underlying meaning of LVEF rather than a simple correlation between the metric and the frames.

Qualitative In Figure 8.4 we show video frame examples to showcase the quality of the recon-

Image	Frames	IFN	MAE	RMSE	R2	Params (M)	FLOPs (G)	Mem (MB)
112×112 [TLYK18a]	32	-	4.22	5.56	0.79	31.3	162	2143.01
112×112[WBBD20]	32	-	5.32	7.23	0.64	346.9	-	12421.0
112×112	64	128	4.32	5.81	0.77	125.1	1280	8472.46
112×112	32	128	4.56	6.16	0.76	125.1	640	4486.52
56×56	64	96	4.71	6.23	0.74	70.4	184	1782.74
56×56	16	128	4.95	6.60	0.71	125.1	82	1000.66
56×56	32	32	5.14	6.70	0.70	7.8	10	281.64
28×28	32	96	5.57	7.17	0.66	70.4	24	471.91
28×28	16	32	7.32	9.25	0.43	7.8	1.3	63.06

Table 8.2: Expert model metrics compared to the model size. We compare results with [TLYK18a] and [WBBD20] as baselines, although the model and sampling methods are different. IFN refers to the number of Initial Feature Maps in the model and defines the number of channels in the entire network. Params (M) is the number of million parameters in the network, FLOPs (G) is the number of billion floating point operations for a forward pass and Mem (MB) is the memory size of a feedforward pass with batch size one.

Model	SSIM	Input Condition	Output
MoCoGAN [TLYK18a]	0.33	Image	Video
ImaGINator [WBBD20]	0.56	Image	Video
D'ARTAGNAN (ours)	0.72	Image	Video

Table 8.3: Images are 64x64, Videos 32 frames, We do not have an expert model in this experiment so we are not evaluating LVEF

structed frames, as well as how the anatomy and style are maintained.

Discussion The predicted LVEF has an MAE of 15.7% which is not optimal. This can come from many factors like the use of the same dataset to train the Expert model and D'ARTAGNAN, the limited number of real videos, or the limited size of the networks due to the necessity of working with three models at once. Those problems could be addressed with hyperparameter search, larger models, as well as additional medical data. For completeness, we compare the performance of D'ARTAGNAN with the literature [WBBD20, TLYK18b], and run additional experiments with an ablated version of our model for conditional video generation, where it achieves the best SSIM score of 0.72.

8.4 Summary

In this chapter we introduced D'ARTAGNAN, a Deep Twin Generative Network able to produce counterfactual images and videos. We showcased its performance in both synthetic and real world medical datasets and achieve visually accurate results and high quantitative scores. Our analysis serves more to exhibit the power of counterfactuals in synthesizing medical images. We admit that the potential actual clinical usage of this method is limited as it does not conform to the usual clinical protocols observed. However, counterfactual inference is still underutilized and under-explored in the field of medical imaging, as such, it is our strong belief that our work can serve as inspiration to future clinically relevant methods.

Chapter 9

Using Causality to Train Machine Learning Algorithms

So far we have been discussing novel approaches to perform the basic tasks of an autonomous diagnostic system. In this chapter we will be taking a step back and investigate how to translate our ML algorithms into the real world. More precisely we will be analyzing under a causal perspective the requirements for extra data. Datasets are both expensive and hard to create in medical images, as such, we need to consider the marginal improvement on our performance if we were going to include more data. This chapter is based upon the "Is more data all you need" paper currently under review. The author was the primary driver of this study by leading the design and implementation of the research

9.1 Introduction

Translating deep learning methods to new applications in the clinic often start with two questions that are very difficult to answer: How much data to collect and what data aspects need more attention than others to meet clinical performance expectations. In diagnostic settings, performance is often characterized by a biased metric, for example an expectation towards zero false negatives so that no signs of disease are missed but some leeway towards false positives,

which can be mitigated with further diagnostic tests. However, this commonly leads to situations where end-users request from machine learning practitioners to make specific interventions on well-performing models, for example to make a deep neural network more sensitive towards one specific class of disease or to change predictions for a selected group of patients, while keeping its sensitivity and specificity intact for other classes.

Active learning [BRK21] is one option to make such interventions on a working model but it has been shown that the introduced bias is not necessarily beneficial and might harm a model's specificity [FGR20]. Furthermore, there is currently no method to estimate for how long further (active) learning should go ahead or how many more samples of specific classes have to be collected until the expected change can be observed. This has critical implications for practical translation of such methods into the clinical practice since time, costs and amount of data cannot be estimated in advance, which in turn conflicts with the need for data minimization as recommended by General Data Protection Regulations [VVdB17].

We believe that methods like ours should be integral parts of regulation approval processes. As models undergo fine-tuning and retraining in the process of development and commercialization we should maintain the same high standards of accuracy and robustness. We need hence to be able to provide guarantees that the resulting ML models cannot degrade in performance.

Therefore, the need arises to be able to reason about the data needs of an application and decide upon the best allocation of resources. Moving forward from the well-known active learning paradigm we are looking at more targeted intervention scenarios and provide in this chapter a causal approach that allows to estimate how much *extra* data is needed for targeted interventions on trained deep learning models. We show on a synthetic dataset and an exemplary large Diabetic Retinopathy (DR) medical imaging dataset how to use our approach.

Contribution: We treat the aforementioned scenarios as a counterfactual meta analysis upon a static model. Our goal is to highlight causal analysis as a potential alternative to active learning and showcase the powerful insights it could yield. Interestingly, we found that it is not always advantageous to increase the size of a dataset.

9.2 Related Work

Recent works on the field of model performance analysis have primarily been focused on determining the required number of samples to achieve. [FZTKN12] developed a inverse power law model to predict model performance with different data sizes. [CLS+15, BNB+13] performed empirical studies on the learning behavior of classifiers to determine sample size requirements. None, however, of the above methods are able to determine the effect of interventions on individual samples, which is the primary concern of this investigation. The closest work is [YXK+21] where the authors have identified the same problem statement of consistency between re-training of algorithms as this work but opt for a re-weighting approach under the name Focal Distillation. Causality, on the other hand, is the field of analysis of causal relationships between variables. The field was expanded to computer science by the works of, among others, J. Pearl [Pea09], however few works exist on the intersection of machine learning, medical imaging and causal analysis. Recently, discussed as useful for medical image analysis [CWG20], More commonly such approaches are found in fields like econometrics [IR15], epidemiology [CK20] and clinical medicine[LSM+17, RLJ20, OS19]. We are borrowing inspiration from these works to argue the potential advantages of causal analysis of medical imaging machine learning algorithms.

9.3 Method

As discussed in Section 2.9 there exists a wide range of possible methods that enable accurate estimation of counterfactual quantities in real world scenarios [SLK18, CWG20, VKGL21]. For the scope of this chapter we are assuming perfect knowledge and as such we revert to the mathematical tools that counterfactual inference methods approximate. We are going to be mainly concerned about the effect of datasets on the classification outcome of the samples. We treat the model architecture and its hyper parameters as confounders that affect only the outcome and are invariant between different treatments of our dataset. We intervene on the size and composition of our dataset leaving all other parameters the same. In Figure 9.1 we

show a sample directed acyclic graph (DAG) for the causal relationships between the variables we take into account for our analysis. For simplicity for each counterfactual we explore only interventions on one of the possible treatments.

Our main research question is a counterfactual one; given a sample was incorrectly classified would this sample be correctly classified if we had trained our model with a different dataset? Mathematically this resembles the probability of Sufficiency $P_S = P(Y_{X=T} = y \mid X = T', Y = y')$ and can be described as:

$$P(Y = 1, do(X = \hat{\mathcal{D}}) \mid Y = 0, X = \mathcal{D}, Z) \tag{9.1}$$

where Y is the outcome of whether or not the sample was correctly classified; X represents the treated dataset $\mathcal{D}$ and Z the confounders like the architecture of the model. Our analysis regards the model as a black box and its underlying architecture complexity is not a constraint. We use this question as an example for a potential causal analysis on a medical image model. We highlight that multiple counterfactual questions can be asked, concerning for example the model architecture or a specific feature of the data. We chose interventions on the overall statistics of the dataset as an entry level counterfactual question with evident real world impact on ML practice.

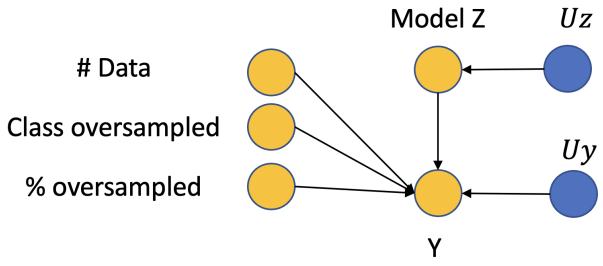


Figure 9.1: A DAG showing the causal relationships between the factors we are analyzing - Yellow: Observed variables - Blue: Unobserved Variables. For simplicity we intervene on one of the possible treatment variables during each experiment. This DAG is a general framework of how we believe the underlying system. To be precise with Model Z we represent the model architecture, as this is set by us a-priori and remains static, the effect of U_z is non existent.

9.4 Evaluation

9.4.1 Datasets

We use two datasets, one synthetic and one real medical imaging database. As synthetic data we use the MorphoMNIST [CTK+19] dataset. This dataset provides a series of morphological operations upon the digits of the well known MNIST dataset [LBBH98]. Each of the generated datasets is modified by random fractures and swellings. In order to control this perturbation in our dataset throughout the experiments each sample retains the same morphological perturbation.

Furthermore, we chose Kaggle's Retinopathy open source dataset [Kag22] as medical imaging dataset. We treat each of the images as a gray scale image and resize it down to $128 \times 128px$. As this is a multi-class dataset with heavy imbalance, we focus on the following labels for Diabetic Retinopathy (DR) levels : "No DR", "Mild" and "Moderate" cases.

9.4.2 Model architecture

As a first step we train a classifier with the full training set. The goal of this step is to determine an architecture and set of hyper-parameters that are able to provide acceptable classification results. Having determined such parameters we fix them for each counterfactual query as to determine the true causal effect of our dataset interventions on the probability of correct classification of a sample. For the synthetic case we opt for a multi-layered Perceptron while for the medical use case a simple multi scale residual convolutional net. While in the synthetic case we opt for a simple model and in the medical one with a significantly more complex one, we note that this is not a parameter that imposes any constraints on our analysis.

9.4.3 Interventions

Interventions are focused on the size and composition of each dataset. First we explore the effect of the size of the dataset. Given that both the synthetic and medical tasks have an abundance of data we create a series of datasets with $[100, 1000, 5000, 10000]$ samples for the synthetic dataset and $[100, 500, 1000, 2000, 4000]$ samples from the medical dataset. We use the intervened datasets to train our models and a static test dataset that includes 20% of the full dataset. This serves as a gold standard evaluation set. Moreover, we intervene upon the dataset by modulating the number of samples in selected classes as well as the percentage of the base dataset by which we will increase a class. In other words given a base dataset of length $N_{dataset}$ with approximate class balance and an upsampling percentage of 10% with class 2, we will add $10\% \times N_{dataset}$ that has unseen samples of class 2. We explore upsampling all available classes by one of the following percentages $0\%, 5\%, 10\%, 20\%, 30\%, 50\%$

9.5 Results

9.5.1 Synthetic Data

First we evaluate the synthetic task. In Table 9.1a we show the effect of including more data regardless of the type of data given a base dataset. We notice for example that if we had 5k samples and switched to a dataset with 10k incorporating randomly selected samples that would give us only a 9% chance of correctly classifying a sample that was previously misclassified. As a general observation we can see that the more samples are contained in our *base dataset* the smaller the probability of a specific sample being correctly classified after the intervention of including more images. A proxy to this insight can be seen in Table 9.1b where we show the average F1 score for all classes for different sized datasets. We see that indeed the overall improvement between 5 and 10 thousand samples is quite small. Meanwhile in Figures 9.2 and 9.3 we show the probability of a sample "flipping" from being misclassified to being correctly classified versus the class we upsample and the percentage of upsampling.

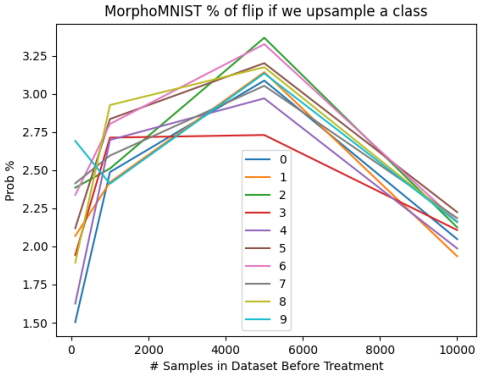


Figure 9.2: Probability of flip; Treatment: Upsampling; Here we upsample a chosen class (seen in legend) and measure the probability of correctly classifying a previously misclassified sample

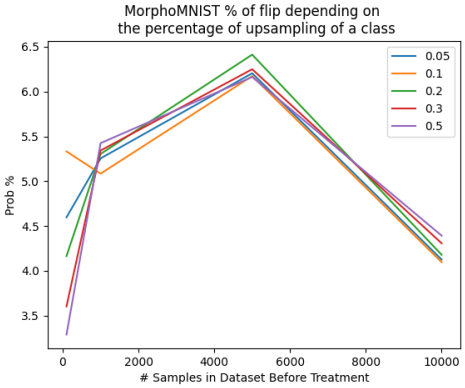


Figure 9.3: Probability of flip; Treatment: Upsampling percentage of base dataset; In this experiment we modulate the level of upsampling a digit

In both cases we look at any given sample without taking into account the ground truth label of it. It is evident from both figures that there is no single class or percentage of upsampling that is key for this dataset – in other words, there is no category of samples that contain key information to help the classification task, rather all classes seem to be necessary.

Table 9.1: MorphoMNIST if we change the number of samples regardless of class and percentage. (A) The probability of changing a misclassified sample to a correctly classified one is shown.(B) F1 performance of different sized base datasets

(a)

# of samples From / To	100	1000	5000	10000
100	0	23.17%	27.18%	27.72%
1000	4.17%	0	17.57%	18.93%
5000	3.95%	7.22%	0	9.46%
10000	3.94%	7.20%	7.25%	0

(b)

	F1
100	30.606%
1000	75.97%
5000	85.54%
10000	86.84%

Thus far we have increased class samples and adding more data to our base dataset without taking into consideration the true class of the misclassified data. In the next step we look into informed interventions where we include a larger number of samples where the majority of them are from the class which was misclassified. In Table 9.2 we show the effects of different dataset sizes and percentages of upsampling on the probability of a specific sample being correctly classified after the intervention. Compared to incorporating more data randomly or

	100	1000	5000	10000
full dataset	7.71	21.18	23.35	24.48
100	0	29.16	30.12	30.32
1000	19.34	0	24.30	24.92
5000	18.70	17.41	0	18.58
10000	18.49	16.94	16.69	0

(a)

	0.05	0.1	0.2	0.3	0.5
full dataset	8.13	9.21	11.49	13.86	15.72
0.05	0	11.50	14.36	17.16	19.61
0.1	9.99	0	14.64	17.21	19.66
0.2	10.64	12.03	0	17.53	19.83
0.3	11.13	12.68	15.27	0	20.02
0.5	11.86	13.27	15.87	18.27	0

(b)

Table 9.2: Informed Interventions (A) the effect of different dataset sizes (B) the effect of different upsampling percentages

upsampling a class regardless of the misclassfied class we observe a significantly increased effect. If, conversely, we do not specify the extent we upsample the misclassified class we average 67.56% probability of correctly classifying a sample after the intervention across all possible percentages and dataset sizes. It is evident, thus, that we can obtain a higher probability of sufficiency if our interventions on the dataset are targeted. We further note that during this analysis we look at the inverse probability of incorrectly classifying a sample that was correctly determined before treatment, in all our cases this probability did not exceed $3-4\%$ indicating that our chosen interventions did not have a negative effect upon the model performance.

9.5.2 Retinopathy

We follow a similar analysis for the medical image data where we classify some of the most abundant categories of the open source Retinopathy dataset. In Table 9.3 we show the two most interesting results. In this real world dataset we observe that incorporating more of the *moderate DR* class leads to all together better classification performance regardless of the dataset size. On the other hand modulating the overall number of samples under an informed sampling regime seems to be driven primarily by our informed sampling than the actual changes in number of data. We note that increasing the datasize from 100 to 2000 by randomly sampling from the available classes only provides a $\sim 10\%$ chance of a sample flipping. Medical imaging datasets can offer interesting insights if looked under a causal prism. It is possible to identify inter-dependencies of classes and features and hence able to plan the dataset acquisition and

annotations more efficiently.

9.6 Discussion

We analyzed the effect of dataset size and composition on the probability of a specific misclassified sample to become correctly classified after our interventions. We have observed a wide expected range for this causal probability. If used in practice to analyse a phenomenon and determine the best allocation of resources, certain thresholds that make sense have to be determined by the users. Contrary to the well-known active learning paradigm where we are interested in the effect of an intervention on the overall metrics of our task, by assuming a causal perspective we are able to estimate the effect of interventions or counterfactuals on an individual sample. This ability, enables a finer grain analysis of our interventions and their effects. For the purposes of our analysis we have assumed complete knowledge of the behavior of our models under different data regimes. This however, is not a valid assumption in real life model development. In such cases, the practitioner could employ a method from literature to estimate or bound the above probability of causation.

If we do not condition on the knowledge that the sample was initially misclassified and we are solely interested in the interventional probability if it will be correctly classified, we aim to learn the conditional average treatment effects: $\mathbb{E}(Y_{X=1} \mid Z) - \mathbb{E}(Y_{X=0} \mid Z)$. Examples of methods that can estimate these include PerfectMatch [SLK18], DragonNet [SBV19], PropensityDropout [AWVDS17], Treatment-agnostic representation networks (TARNET) [SJS16], Balancing Neural Networks [JSS16]. Other machine learning approaches to estimating interventional queries made use of GANs, such as GANITE [YJVDS18] and CausalGAN [KSDV18], Gaussian Processes [WTJM20, AvdS17], Variational Autoencoders [LSM$^+$17], and representation learning [ZBS20, AZT$^+$21, YLL$^+$18].

If, on the other hand, we wish to answer the same query in Equation (9.1) we need to utilize methods able to handle counterfactual queries. Recent work [PCG20] proposes normalizing flows and variational inference to compute counterfactual queries using abduction-

	Healthy	Mild	Moderate
Healthy	0	12.03	18.75
Mild	12.95	0	14.74
Moderate	16.52	11.31	0

(a)

	100	500	1000	2000
100	0	18.28	18.16	19.77
500	18.23	0	18.00	19.70
1000	18.33	18.04	0	18.79
2000	18.88	18.71	17.59	0

(b)

Table 9.3: Dataset Interventions on the Retina dataset (A) the effect of different up-sampling classes (B) the effect informed interventions and dataset sizes

action-prediction. [OS19] used the Gumbel-Max trick to estimate counterfactuals, again using abduction-action-prediction. While this methodology satisfies generalizations of identifiability constraints. Additional work by [CK20] devised a non-parametric method to compute the Probability of Necessity using an influence-function-based estimator. A limitation of this approach is that a separate estimator must be derived and trained for each counterfactual query. Finally, [VKGL21] estimate counterfactual probabilities via means of a deep twin network while imposing identifiability constraints in the case of binary treatments and outcomes.

Finally, we should comment on the relation of our causal methods when compared with other "classical" ML algorithms. While other methods might be able to estimate accurately the class of samples that would decrease the average uncertainty statistics (active learning), there are no guarantees that the underlying causal relations have been identified. As such the same method might not be able to correctly identify the required samples if the regime has undergone even slight changes. Moreover, we should point out that we do not assume that performance of the model has reached or will reach a plateau with out methods. It is commonly believed that modern ML algorithms seem not to be reaching a plateau in performance. In summary we position our causal framework as a low-shot active learning regime that can minimize the number of samples needed to re-train and ameliorate the model.

9.7 Summary

With the work detailed in this chapter we hoped to stimulate a new topic of discussion in the ML and Medical imaging community around causal analysis and how it can help us optimize resource allocations. Being able to quantify the per sample effect of an intervention is necessary to better understand a given task. Besides economical constraints, the proven environmental impact of our field [Dha20] means that we cannot opt for increasingly larger models when the expected returns are minimal. Causally analyzing the task at hand can provide estimates of performance vs. computational and economical resources without the need to run the experiments.

Chapter 10

Conclusions

10.1 Summary of Thesis Achievements

Throughout the thesis we have been concerned with a major objective: How to build an autonomous diagnostic system that is robust enough to work in new and demanding environments as diverse as deep space exploration missions. We have looked into exploring and identifying points of medical interest. We have developed methods to imagine how objects look like in 3D after being observed in 2D, and recall similar information from memory. Finally we have developed a series of novel algorithms to enable the agents to causally reason about the information they have extracted and predict future outcomes of their actions.

In **Chapter 3** we lay out a case study relating to deep space exploration. We argue that as humans venture out to Mars and the rest of the solar system we cannot rely on real time communication with doctors on Earth and telemedicine as the communication delay is prohibitively large. Moreover due to weight there are limitations on personnel and even the astronauts on board cannot have expertise in all possible scenarios. As such we highlighted the need for a robust, causally enabled, autonomous diagnostic system.

Chapter 4 commences the technical innovations of this thesis. We proposed a method that allows multiple agents explore their respective environments and locate medical landmarks.

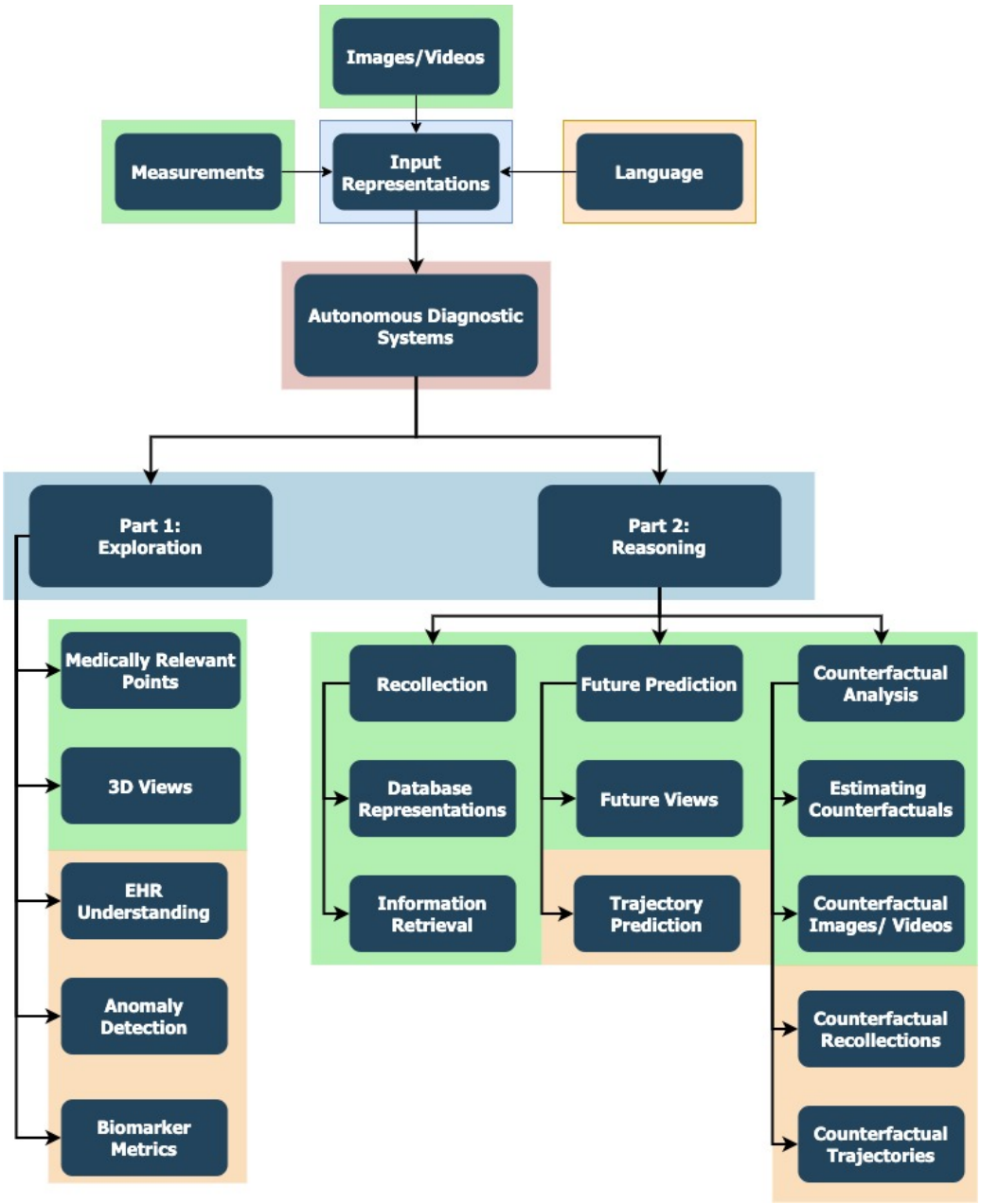


Figure 10.1: Reminder overview of key elements of a hypothetical Autonomous Diagnostic System

Our rationale stems from domain knowledge that anatomical landmarks are interdependent in their locations in space - in other words knowing where one is can help locate the others. By sharing weights in the backbones of the ML system we are able to leverage diverse information sources that can inform the model of the aforementioned anatomical interdependencies. We showcased our results in two medical modalities - MRI and US - beating the state of the art in RL anatomical landmark detection. While subsequent works have focused in non-RL methods to identify anatomical landmarks, the need for guidance in finding the landmarks call for the sequential decision making that RL offers.

Chapter 5 builds upon another required element of a truly automated diagnostic system - 3D understanding of objects. As medical images are 2D depictions of 3D objects; being able to understand how the depicted shape behaves in the real 3D world is very important for reasoning and challenging as its an ill-defined problem. Recognizing the multitude of possible solutions we assume a probabilistic approach and augment the Phi-Seg model from literature to be able to do 2D to 3D reconstruction. Besides employing 3D convolutions we also introduce a structural reconstruction module which transforms 2D images into 3D representations that are then fed into the 3D Phi-Seg model. Moreover we employ a fusion module that fuses the 2D DRR image's texture and shape information with the reconstructed 3D volume. Finally we test our method in DRR to CT reconstruction and X-Ray segmentation of human and porcine lungs and rib cages. Since the development of this work we have seen a significant advance via the use of Neural Radience Fields (NERF) and diffusion models which currently set the state of the art.

While in Part I we were discussing how to extract information from observations and perform various tasks like reconstruction, segmentation, localization and information retrieval; in Part II we will be leveraging the extracted information and perform causally enabled reasoning tasks.

Chapter 6 showcases one potential application of causal ML in the task of future frame prediction. In a deviation from the norm of ML methodology we combine concepts from general relativity - the notion of the light cone - to restrict our learned pseudo-Riemannian latent space where from we sample. Showing promising results in synthetic and real life settings we also lay

out a methodology on how to perform causal inference under both Pearl's SCM and Rubin's Potential Outcomes frameworks.

Chapter 7 develops a novel counterfactual inference technique based on Balke & Pearl's [BP94] Twin Networks. We combine our Deep Twin Networks with identifiability constraints for binary outcome-treatment pairs and derive novel theoretical results in form of extending the aforementioned constrain to categorical variables - we call this constraint Counterfactual Ordering. We exhibited experimentally the validity of our theoretical results as well as the computational efficiency of the Deep Twin Networks.

In **Chapter 8** we extended the Deep Twin Networks to medical imaging. In this scenario we had to tackle two main causal inference challenges. High dimensional inputs create an array of Causal Markov Equivalent models that we are not able to identify easily from observational data. Moreover the lack of true counterfactual targets poses a significant challenge for supervised algorithms like ours. In order to resolve these issues we reduce the dimensionality of the data via a VQ-VAE and perform the causal inference in the lower dimensional manifold. Moreover we employ an expert network to regress the LVEF from the counterfactual outcome, as well as a discriminator to lead the texture and image quality training. We evaluated our methods in a synthetic task of Morpho-MNIST and a real life cardiac US video sequence dataset.

Finally, **Chapter 9** takes a step back and investigates the use of causal reasoning in developing ML solutions. We showcase a potential application for Business Intelligence and Product Development by assessing the probability of Necessity and the probability of Sufficiency for including more data as it relates to improving model performance. We show that after a certain point the marginal benefit of including more data are not justifying the cost that would incur to the researcher or company.

10.2 Future Work

The work presented in this thesis, the author believes, opens the path for more exciting research works. These range from a complete probabilistic field theory that exists on latent spaces to

applications using causal inference and discovery in the medical and medical imaging domain. I hope that with the works during this PhD that also appear in the thesis, I have inspired and laid the ground works for future researchers to push the limits of human ingenuity and knowledge.

Domain Expertise: As seen in many of the previous chapters the input of domain knowledge and expertise was crucial for achieving the performance observed. Domain knowledge leverages the vastness of human knowledge, intuition and ingenuity to inform the methodological decisions. For Chapter 4 we used the insight that the anatomical landmarks we aim to locate are part of a wider anatomy hence knowing where one is can lead us to another. More insights like the orientation of the landmarks might be also beneficial to the task in future work. Domain knowledge is also crucial in Part II in order to estimate the underlying causal graph. In causal inference we assume that our causal model is known *a-priory*, while in causal discovery we aim to derive the model from observations. In both cases, however, incorporating domain knowledge on how the causal links might be structured or what potential spurious correlations might exist can inform us and help avoid such systematic errors.

Scaling architectures Direct future work that concerns Chapter 6 is the scaling the architectures used. Key limitations of that work were due to the simple proof-of-concept approach adopted. We are strongly believe that deeper convolutional or diffusion based architectures can increase the observed performance. We note that all chapters of this thesis that concern generative methods would benefit from the incorporation of diffusion models that are the current state of the art, but were not known at the time of writing of the original works.

Uses of Causality In Part II we discussed a series of technical and theoretical methods and results that enable causal inference in a wide variety of fields. We believe that future works directly stemming from these advances could include the application of the Deep Twin Networks in *out-of-distribution* and anomaly detection as well as the parametrization of domain adaptation as a causal intervention. We believe that in all these cases, more in-depth analyzed in Section 2.8 and Chapter 9, causal inference can help the community achieve significant advances.

Moreover looking back to Figure 10.1 and the high level diagram of a hypothetical ADS we see a few blocks that will be needed in any effort to build a fully functioning Autonomous Diagnostic System. In this thesis we did not use any textual language based inputs. Language, however, carries a significant amount of information be that through Electronic Health Records or via interactions with the patients. Moreover we did not perform any anomaly detection or biomarker measurements. Both these activities are vital in acquiring more information and establishing a complete picture of the patients state at any given time. Moreover, a crucial component of reasoning about potential diagnosis and treatment plans requires the ability to predict possible and counterfactual trajectories of the patients bio-signals and outcomes. In this way the system can eliminate both potential diagnosis and also assist the attending physician select the appropriate treatment plan.

Finally the work of Chapter 9 can be considered as an introductory preparatory work. While other methods provide argumentation in favor of using causality to enable robust and explainable ML algorithms we showcase the use of causality and causal analysis in a business intelligence task. We believe that future work can include methodological contributions as causally enabled methods to provide estimates of the amount and distributions of data in a dataset. Moreover, we aim at working together with regulators to set out the appropriate thresholds of confidence such that our analysis can inform regulatory procedures, ensuring that a given model is not going to be degraded upon retraining or fine-tuning with data that abide by the identified causal relationships.

Acronyms

A3C Asynchronous Advantage Actor Critic. 58, 156

AC Anterior Commissure. 5, 50, 56, 57, 156

ADS Autonomous Diagnostic System. xxii, xxv, 2, 4, 41, 50, 51, 60, 74, 75, 96, 152, 156

AP Apex. 57, 156

Colab-DQN Collaborative Deep Q-Network. xxii, 54, 55, 56, 57, 59, 156

ConvNet Convolutional neural network. 15, 52, 56, 58, 156

CSP Cavum Septum Pellucidum. xvii, 56, 57, 58, 156

CT Computed Tomography. 4, 12, 44, 60, 61, 62, 63, 68, 70, 71, 72, 156

CXR Chest X-Ray. 61, 156

D'ARTAGNAN Deep ARtificial Twin-Architecture GeNerAtive Networks. xxiv, 130, 133, 135, 136, 137, 138, 139, 156

DL Deep Learning. 6, 156

DQN Deep Q-Network. xvii, 51, 52, 55, 56, 57, 58, 59, 156

DRR Digitally Reconstructed Radiographs. 60, 62, 66, 67, 68, 70, 72, 156

ED End-Diastolic. xxiv, 134, 137, 156

ES End-Systolic. xxiv, 134, 137, 156

GAN Generative Adversarial Network. 62, 63, 130, 132, 136, 156

GANs Generative Adversarial Networks. 156

IR Information Retrieval. 156

LC Left Cerebellum. 57, 58, 156

LEO Lower Earth Orbit. 41, 42, 43, 156

LVEF Left Ventricular Ejection Fraction. xxiv, 129, 132, 134, 135, 136, 137, 138, 154, 156

MARL Multi Agent Reinforcement Learning. 51, 156

MDP Markov Decision Process. 53, 156

ML Machine Learning. 6, 42, 43, 44, 46, 47, 48, 156

MRI Magnetic Resonance Imaging. xvii, 4, 5, 11, 12, 13, 44, 55, 56, 57, 59, 153, 156

MST Minkowski Space-Time. 22, 75, 76, 78, 80, 82, 83, 88, 94, 156

MV Mitral Valve Centre. 57, 156

NN Artificial Neural Network. 13, 156

PC Posterior Commissure. 5, 50, 56, 57, 156

PET Positron Emission Tomography. 11, 12, 156

POCUS Point Of Care Ultrasound. 44, 45, 47, 48, 156

POMDP Partially Observable Markov Decision Process. 53, 156

RC Right Cerebellum. 57, 58, 156

RL Reinforcement Learning. xxii, 45, 48, 51, 52, 53, 56, 156

ROI Region of Interest. 18, 53, 55, 156

SSIM Structural Similarity Metric. xxiii, 92, 156

SV Systolic Volume. 156

TDA Topological Data Analysis. 156

TRL Technology Readiness Level. 42, 156

US Ultrasound. 4, 6, 11, 12, 13, 44, 55, 57, 59, 129, 130, 132, 135, 136, 153, 154, 156

VQ-VAE Vector Quantized-Variational AutoEncoder. 134, 154, 156

VR *Vietoris–Rips. 156*

Bibliography

[AASK+20] Sina Amirrajab, Samaneh Abbasi-Sureshjani, Yasmina Al Khalil, Cristian Lorenz, Juergen Weese, Josien Pluim, and Marcel Breeuwer. Xcat-gan for synthesizing 3d consistent labeled cardiac mr images on anatomically variable xcat phantoms. 7 2020.

[AAUS+19] Muhammad Ayaz, Mohammad Ammad-Uddin, Zubair Sharif, Ali Mansour, and El-Hadi M. Aggoune. Internet-of-things (iot)-based smart agriculture: Toward making the fields talk. *IEEE Access*, 7:129551–129583, 2019.

[AFN17] Shadi Albarqouni, Javad Fotouhi, and Nassir Navab. X-ray in-depth decomposition: Revealing the latent structures. In *MICCAI*, 2017.

[AHH18] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

[AL19] Judea Pearl Ang Li. Unit selection based on counterfactual logic. 2019.

[ALFV+18] Amir Alansary, Loic Le Folgoc, Ghislain Vaillant, Ozan Oktay, Yuanwei Li, Wenjia Bai, Jonathan Passerat-Palmbach, Ricardo Guerrero, Konstantinos Kamnitsas, Benjamin Hou, Steven McDonagh, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Automatic View Planning with Multi-scale Deep Reinforcement Learning Agents. In *MICCAI 18*, pages 277–285, 2018.

[AMAK22] Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *International Conference on Learning Representations*, 2022.

[AOL+19a] Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Konstantinos Kamnitsas, Athanasios Vlontzos, Ben Glocker, Bernhard Kainz, and Daniel Rueckert. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical Image Analysis*, 53:156–164, 2019.

[AOL+19b] Amir Alansary, Ozan Oktay, Yuanwei Li, Loic Le Folgoc, Benjamin Hou, Ghislain Vaillant, Konstantinos Kamnitsas, Athanasios Vlontzos, Ben Glocker, Bernhard Kainz, et al. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical image analysis*, 53:156–164, 2019.

[ASAL+20] Samaneh Abbasi-Sureshjani, Sina Amirrajab, Cristian Lorenz, Juergen Weese, Josien Pluim, and Marcel Breeuwer. 4d semantic cardiac magnetic resonance image synthesis on xcat anatomical model. 2 2020.

[AvdS17] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3427–3435, 2017.

[AVI+16] B. Aubert, C. Vergari, B. Ilharreborde, A. Courvoisier, and W. Skalli. 3d reconstruction of rib cage geometry from biplanar radiographs using a statistical parametric model approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 4(5):281–295, 2016.

[AWVDS17] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966*, 2017.

[AZT+21] Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin Duke. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pages 1972–1980. PMLR, 2021.

[BBH+21] Samuel Budd, Arno Blaas, Adrienne Hoarfrost, Kia Khezeli, Krittika D'Silva, Frank Soboczenski, Graham Mackintosh, Nicholas Chia, and John Kalantari. Prototyping crisp: A causal relation and inference search platform applied to colorectal cancer data. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 517–521. IEEE, 2021.

[BBRH12] Benny Burger, Sascha Bettinghausen, Matthias Radle, and Jürgen Hesser. Real-time gpu-based ultrasound simulation using deformable mesh models. *IEEE transactions on medical imaging*, 32(3):609–618, 2012.

[BCII20] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. Technical report, Columbia University, Stanford University, 2020.

[BGL+93] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. pages 737–744, 1993.

[BJG22] Melanie Bernhardt, Charles Jones, and Ben Glocker. Investigating underdiagnosis of ai algorithms in the presence of multiple sources of dataset bias. *arXiv preprint arXiv:2201.07856*, 2022.

[BNB+13] Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models. *Analytica chimica acta*, 760:25–33, 2013.

[BP94] Alexander Balke and Judea Pearl. Probabilistic evaluation of counterfactual queries. In *AAAI*, 1994.

[BP97] Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

[BPG+21] Mehri Baniasadi, Mikkel V Petersen, Jorge Goncalves, Andreas Horn, Vanja Vlasov, Frank Hertel, and Andreas Husch. Dbsegment: Fast and robust segmentation of deep brain structures–evaluation of transportability across acquisition domains. *arXiv preprint arXiv:2110.09473*, 2021.

[BPP+21] Oualid Benkarim, Casey Paquola, Bo-yong Park, Valeria Kebets, Seok-Jun Hong, Reinder Vos de Wael, Shaoshi Zhang, BT Thomas Yeo, Michael Eickenberg, Tian Ge, et al. The cost of untracked diversity in brain-imaging prediction. *bioRxiv*, 2021.

[BRK21] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 2021.

[BSK+19] Samuel Budd, Matthew Sinclair, Bishesh Khanal, Jacqueline Matthew, David Lloyd, Alberto Gomez, Nicolas Toussaint, Emma Robinson, and Bernhard Kainz. Confident head circumference measurement from ultrasound with real-time feedback for sonographers. *MICCAI*, 2019.

[BTC+19] Christian F. Baumgartner, Kerem C. Tezcan, Krishna Chaitanya, Andreas M. Hötker, Urs J. Muehlematter, Khoschy Schawkat, Anton S. Becker, Olivio Donati, and Ender Konukoglu. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. *MICCAI*, pages 1–14, 2019.

[BUvM+17] Natalia Z Bielczyk, Sebo Uithol, Tim van Mourik, Martha N Havenith, Paul Anderson, Jeffrey C Glennon, and KJ Buitelaar. Causal inference in functional magnetic resonance imaging. *arXiv preprint arXiv:1708.04020*, 2017.

[BWZ+18] Lars Buesing, Theophane Weber, Yori Zwols, Sebastien Racaniere, Arthur Guez, Jean-Baptiste Lespiau, and Nicolas Heess. Woulda, coulda, shoulda:

Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

[ÇAL+16] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sebastien Ourselin and et al., editors, *MICCAI*, 2016.

[Car97] Sean M. Carroll. An introduction to general relativity: spacetime and geometry, 1997.

[CBTRDE21] Jose A Cortes-Briones, Nicolas I Tapia-Rivas, Deepak Cyril D'Souza, and Pablo A Estevez. Going deep into schizophrenia with artificial intelligence. *Schizophrenia Research*, 2021.

[CCL+21] Richard J Chen, Tiffany Y Chen, Jana Lipkova, Judy J Wang, Drew FK Williamson, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithm fairness in ai for medicine and healthcare. *arXiv preprint arXiv:2110.00603*, 2021.

[CFL+22] Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. Neural score matching for high-dimensional causal inference. *AISTATS*, 2022.

[CFS20] Neil J. Cronin, Taija Finni, and Olivier Seynnes. Using deep learning to generate synthetic b-mode musculoskeletal ultrasound images. *Computer Methods and Programs in Biomedicine*, 196:105583, 2020.

[Chi03] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null):507–554, mar 2003.

[CK20] Maria Cuellar and Edward H Kennedy. A non-parametric projection-based estimator for the probability of causation, with application to water sanitation in kenya. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1793–1818, 2020.

[CLD+21] Xiaoyang Chen, Chunfeng Lian, Hannah H. Deng, Tianshu Kuang, Hung-Ying Lin, Deqiang Xiao, Jaime Gateno, Dinggang Shen, James J. Xia, and Pew-Thian Yap. Fast and accurate craniomaxillofacial landmark detection via 3d faster r-cnn. *IEEE Transactions on Medical Imaging*, 40(12):3867–3878, 2021.

[CLS+15] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.

[COL+11] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158–e177, 2011.

[CRBC22] Kai-Cheng Chuang, Sreekrishna Ramakrishnapillai, Lydia Bazzano, and Owen Carmichael. Nonlinear conditional time-varying granger causality of task fmri via deep stacking networks and adaptive convolutional kernels. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 271–281, Cham, 2022. Springer Nature Switzerland.

[CTK+19] Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-MNIST: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20(178), 2019.

[CWG20] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, Jul 2020.

[CXL+] Victor Manuel Campello, Tian Xia, Xiao Liu, Pedro Sanchez, Carlos Martín-Isla, Steffen Erhard Petersen, Santi Seguí, Sotirios Tsaftaris, and Karim Lekadir. Cardiac aging synthesis from cross-sectional data with conditional

generative adversarial networks. *Frontiers in cardiovascular medicine*, page 2693.

[CYLW13] Weijian Cong, Jian Yang, Yue Liu, and Yongtian Wang. Fast and automatic ultrasound simulation from ct images. *Computational and mathematical methods in medicine*, 2013, 2013.

[Dem17] Robert C. Dempsey. *The International Space Station: operating an outpost in the new frontier*. National Aeronautics and Space Administration, Lyndon B. Johnson Space Center, 2017.

[DF18] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1174–1183, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[DG16] Jayaraman Dinesh and Kristen Grauman. Look-Ahead Before You Leap : End-to-End Active Recognition. *ECCV*, 2016.

[DG17] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Dha20] Payal Dhar. The carbon impact of artificial intelligence. *Nature Machine Intelligence*, 2, 2020.

[Die00] Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

[DJL20] Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *medRxiv*, 2020.

[dMDS+14] Antonio de Marvao, Timothy JW Dawes, Wenzhe Shi, Christopher Minas, Niall G Keenan, Tamara Diamond, Giuliana Durighel, Giovanni Montana,

Daniel Rueckert, Stuart A Cook, et al. Population-based studies of myocardial hypertrophy: high resolution cardiovascular magnetic resonance atlases improve statistical power. *Journal of cardiovascular magnetic resonance*, 16(1):16, 2014.

[DSDB16] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[dSGS+20] Mariana da Silva, Kara Garcia, Carole H Sudre, Cher Bass, M Jorge Cardoso, and Emma Robinson. Biomechanical modelling of brain atrophy through deep learning. *arXiv preprint arXiv:2012.07596*, 2020.

[DV16] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[DWX+20] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z Jane Wang. Ccgan: Continuous conditional generative adversarial networks for image generation. In *International Conference on Learning Representations*, 2020.

[DZK+22] Hao Ding, Jintan Zhang, Peter Kazanzides, Jie Ying Wu, and Mathias Unberath. Carts: Causality-driven robot tool segmentation from vision and kinematics data. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 387–398, Cham, 2022. Springer Nature Switzerland.

[Ein15] Albert Einstein. *Relativity.* Princeton University Press, 2015.

[EJRB+18] S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Theophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil Rabinowitz, Helen King, Chloe Hillier, Matt Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[FAdFW16] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *NIPS 29*, pages 2137–2145, 2016.

[FCAS+18] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proc. 17th Intl. Conf. on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pages 122–130, 2018.

[FDA21] FDA. Medical x-ray imaging. https://www.fda.gov/radiation-emitting-products/medical-imaging/medical-x-ray-imaging, 2021.

[FGR20] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2020.

[FPKK22] Jana Fehr, Marco Piccininni, Tobias Kurth, and Stefan Konigorski. A causal framework for assessing the transportability of clinical prediction models. *medRxiv*, 2022.

[FZTKN12] Rosa L Figueroa, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. Predicting sample size required for classification performance. *BMC medical informatics and decision making*, 12(1):1–10, 2012.

[GBH18] Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic Entailment Cones for Learning Hierarchical Embeddings. *ICML*, 2018.

[GCC+09] Hang Gao, Hon Fai Choi, Piet Claus, Steven Boonen, Siegfried Jaecques, G Harry Van Lenthe, Georges Van der Perre, Walter Lauriks, and Jan

D'hooge. A fast convolution-based methodology to simulate 2-dd/3-d cardiac ultrasound images. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 56(2):404–409, 2009.

[GCLB15] Romane Gauriau, Rémi Cuingnet, David Lesage, and Isabelle Bloch. Multi-organ localization with cascaded global-to-local regression and shape prior. *Medical Image Analysis*, 23(1):70 – 83, 2015.

[GCP+16] Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander Van Esbroeck. Monotonic calibrated interpolated look-up tables. *The Journal of Machine Learning Research*, 17(1):3790–3836, 2016.

[GE15] Justin Girard and Reza Emami. Concurrent markov decision processes for robot team learning. *EAAI*, 2015.

[GEH21] GEHealthcare. Point of care ultrasounds. https://www.gehealthcare.co.uk/products/ultrasound/point-of-care-ultrasound, 2021.

[GEK17] Jayesh K. Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems*, pages 66–83. Springer, 2017.

[GG16] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *ICML*, 2016.

[GGCV95] Davi Geiger, Alok Gupta, Luiz A. Costa, and John Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(3):294–302, 1995.

[GGLT21] Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. A counterfactual simulation model of causal judgments for physical events. *Psychological review*, 128(5):936, 2021.

[GGM+16] Florin C. Ghesu, Bogdan Georgescu, Tommaso Mansi, Dominik Neumann, Joachim Hornegger, and Dorin Comaniciu. An artificial agent for anatomical landmark detection in medical images. In *MICCAI 2016*, pages 229–237, Cham, 2016. Springer.

[GGZ+19] F. Ghesu, B. Georgescu, Y. Zheng, S. Grbic, A. Maier, J. Hornegger, and D. Comaniciu. Multi-scale deep reinforcement learning for real-time 3d-landmark detection in ct scans. *IEEE PAMI*, 41(1):176–189, Jan 2019.

[Gha20] Nilesh Gharia. Dnn radiation hardened co-processor companion chip to nasa's upcoming high-performance spaceflight computing processor. https://www.sbir.gov/node/1882269, 2020.

[GKC+17] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, David Lopez-Paz, Isabelle Guyon, Michele Sebag, Aris Tritas, and Paola Tubaro. Learning functional causal models with generative neural networks. *arXiv preprint arXiv:1709.05321*, 2017.

[GKC+18] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, pages 39–80. Springer, 2018.

[GKGLF20] Daniel Grzech, Bernhard Kainz, Ben Glocker, and Loïc Le Folgoc. Image registration via stochastic gradient markov chain monte carlo. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pages 3–12. Springer, 2020.

[GLA19] Daniele Grattarola, Lorenzo Livi, and Cesare Alippi. Adversarial autoencoders with constant-curvature latent manifolds. *Applied Soft Computing*, 2019.

[GLP19] Logan Graham, Ciarán M Lee, and Yura Perov. Copy, paste, infer: A robust analysis of twin networks for counterfactual inference. *NeurIPS Causal*

ML workshop 2019, https://cpb-us-w2.wpmucdn.com/sites.coecis.cornell.edu/dist/a/238/files/2019/12/Id_65_final.pdf, 2019.

[GMD+20] Tim Genewein, Tom McGrath, Grégoire Déletang, Vladimir Mikulik, Miljan Martic, Shane Legg, and Pedro A. Ortega. Algorithms for Causal Reasoning in Probability Trees. oct 2020.

[GML21] Fabio Garcea, Lia Morra, and Fabrizio Lamberti. On the use of causal models to build better datasets. In *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1514–1519. IEEE, 2021.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GPS89] Dorothy M Greig, Bruce T Porteous, and Allan H Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51(2):271–279, 1989.

[GPS21] Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. *arXiv preprint arXiv:2103.11972*, 2021.

[GSCHH] Beatriz Garcia Santa Cruz, Andreas Husch, and Frank Hertel. The effect of dataset confounding on predictions of deep neural networks for medical imaging.

[GVMB22] Pedro M Gordaliza, Juan José Vaquero, and Arrate Munoz-Barrutia. Translational lung imaging analysis through disentangled representations. *arXiv preprint arXiv:2203.01668*, 2022.

[GZS19] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[HAT+20] Hannes Hase, Mohammad Farid Azampour, Maria Tirindelli, Magdalini Paschali, Walter Simson, Emad Fatemizadeh, and Nassir Navab. Ultrasound-guided robotic navigation with deep reinforcement learning, 2020.

[HDES+22] Andreas Holzinger, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, Wojciech Samek, Igor Jurisica, and Natalia Díaz-Rodríguez. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion*, 79:263–278, 2022.

[HGDG17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[HIP21] HIPPA. Hippa data protection guidelines. https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations/index.html, 2021.

[HJL+20] D. M. Hiemstra, C. Jin, Z. Li, R. Chen, S. Shi, and L. Chen. Single event effect evaluation of the jetson agx xavier module using proton irradiation. In *2020 IEEE Radiation Effects Data Workshop (in conjunction with 2020 NSREC)*, pages 1–4, 2020.

[HKY20] Grant Haskins, Uwe Kruger, and Pingkun Yan. Deep learning in medical image registration: a survey. *Machine Vision and Applications*, 31(1):1–18, 2020.

[HMT21] Hokuto Hirano, Akinori Minagi, and Kazuhiro Takemoto. Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging*, 21(1):1–13, 2021.

[HRRR18] Philipp Henzler, Volker Rasche, Timo Ropinski, and Tobias Ritschel. Single-image tomography: 3d volumes from 2d cranial x-rays. *Comput. Graph. Forum*, 37(2):377–388, 2018.

[HWB+19] Yixing Huang, Tobias Würfl, Katharina Breininger, Ling Liu, Günter Lauritsch, and Andreas Maier. Abstract: Some investigations on robustness of deep learning in limited angle tomography. In Heinz Handels, Thomas M. Deserno, Andreas Maier, Klaus Hermann Maier-Hein, Christoph Palm, and Thomas Tolxdorff, editors, *MICCAI*, pages 21–21, Wiesbaden, 2019. Springer Fachmedien Wiesbaden.

[HZZ+20] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data with independent changes. *JMLR*, Jun 2020.

[IA94] Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

[IR15] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[JEEL19] Dinesh Jayaraman, Frederik Ebert, Alexei Efros, and Sergey Levine. Time-agnostic prediction: Predicting predictable video frames. In *International Conference on Learning Representations*, 2019.

[Jia10] Wu Jian. ITK-Based Implementation of Two-Projection 2D/3D Registration Method with an Application in Patient Setup for External Beam Radiotherapy. In *Insight Journal*, 2010.

[JJ+08] Clifford R Jack Jr et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[JLH+18] Rong Jiao, Nan Lin, Zixin Hu, David A. Bennett, Li Jin, and Momiao Xiong. Bivariate causal discovery and its applications to gene expression and imaging data analysis. *Frontiers in Genetics*, 9, 2018.

[JSJ95] Tommi Jaakkola, Satinder P. Singh, and Michael I Jordan. Reinforcement learning algorithm for partially observable markov decision problems. *NIPS*, 1995.

[JSS16] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pages 3020–3029, 2016.

[JZ21] Elias Bareinboim Junzhe Zhang, Jin Tian. Partial counterfactual identification from observational and experimental data. 2021.

[Kag22] Kaggle. Diabetic retinopathy detection dataset. *Kaggle.com*, 2022.

[Kar16] Andrej Karpathy. Cs231n convolutional neural networks for visual recognition. *Stanford Computer Science; CS231n*, 2016.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KCW+22] Nan Rosemary Ke, Silvia Chiappa, Jane Wang, Jorg Bornschein, Theophane Weber, Anirudh Goyal, Matthew Botvinic, Michael Mozer, and Danilo Jimenez Rezende. Learning to induce causal structure. 2022.

[KHN+22] Amar Kumar, Anjun Hu, Brennan Nichyporuk, Jean-Pierre R Falet, Douglas L Arnold, Sotirios Tsaftaris, and Tal Arbel. Counterfactual image synthesis for discovery of personalized predictive image markers. *arXiv preprint arXiv:2208.02311*, 2022.

[KK14] Olga Kosheleva and Vladik Kreinovich. Observable causality implies lorentz group: alexandrov-zeeman-type theorem for space-time regions. *Mathematical Structures and Modeling*, 2014.

[KLMP15] Michael Kremer, Jessica Leino, Edward Miguel, and Alix Peterson. Replication data for: Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions. 2015.

[KLN+17] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[KLRS17] MJ Kusner, J Loftus, Christopher Russell, and R Silva. Counterfactual fairness. *Advances in Neural Information Processing Systems 30 (NIPS 2017) pre-proceedings*, 30, 2017.

[KMY+16] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[KØP+19] Wouter M. Kouw, Silas N. Ørting, Jens Petersen, Kim S. Pedersen, and Marleen de Bruijne. A cross-center smoothness prior for variational bayesian brain tissue segmentation. In Albert C. S. Chung, James C. Gee, Paul A. Yushkevich, and Siqi Bao, editors, *Information Processing in Medical Imaging*, pages 360–371, Cham, 2019. Springer International Publishing.

[KPB12] Thomas Kroes, Frits H Post, and Charl P Botha. Exposure render: An interactive photo-realistic volume rendering framework. *PloS one*, 7(7), 2012.

[KSDV18] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

[KSMZ+20] Maxime Kayser, Roger D Soberanis-Mukul, Anna-Maria Zvereva, Peter Klare, Nassir Navab, and Shadi Albarqouni. Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting. *arXiv preprint arXiv:2002.02883*, 2020.

[KTY+18] Thanard Kurutach, Aviv Tamar, Ge Yang, Stuart Russell, and Pieter Abbeel. Learning plannable representations with causal infogan. *Advances in Neural Information Processing Systems*, 2018.

[KVSe14] Bernhard Kainz, Philip Voglreiter, Michael Sereinigg, and et al. High-Resolution Contrast Enhanced Multi-Phase Hepatic Computed Tomography Data from a Porcine Radio-Frequency Ablation Study. 2014.

[KW09] Christopher Koehler and Thomas Wischgoll. Knowledge-assisted reconstruction of the human rib cage and lungs. *IEEE computer graphics and applications*, 30(1):17–29, 2009.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[LAC+18] Yuanwei Li, Amir Alansary, Juan Cerrolaza, Bishesh Khanal, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Fast multiple landmark localisation using a patch-based iterative network. In *Proceedings 21st International Conference, Granada, Spain, September 16–20, 2018, Part I*, pages 563–571, Sep. 2018.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LCKD+05] María J Ledesma-Carbayo, Jan Kybic, Manuel Desco, Andrés Santos, Michael Suhling, Patrick Hunziker, and Michael Unser. Spatio-temporal nonrigid registration for ultrasound cardiac motion estimation. *IEEE transactions on medical imaging*, 24(9):1113–1126, 2005.

[LGLV+21a] Alexander Lavin, Ciarán M Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Sujoy Ganguly, Danny Lange, Atılım Güneş Baydin, Amit Sharma, Adam Gibson, et al. Technology readiness levels for machine learning systems. *arXiv preprint arXiv:2101.03989*, 2021.

[LGLV+21b] Alexander Lavin, Ciarán M. Gilligan-Lee, Alessya Visnjic, Siddha Ganju, Dava Newman, Sujoy Ganguly, Danny Lange, Atılım Güneş Baydin, Amit Sharma, Adam Gibson, Yarin Gal, Eric P. Xing, Chris Mattmann, and James Parr. Technology readiness levels for machine learning systems, 2021.

[LGZ13] David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073, 2013.

[LJM+21] Guy Lorberbom, Daniel D Johnson, Chris J Maddison, Daniel Tarlow, and Tamir Hazan. Learning generalized gumbel-max causal mechanisms. *Advances in Neural Information Processing Systems*, 34:26792–26803, 2021.

[LKL18] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[LLLG18] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future Frame Prediction for Anomaly Detection—A New Baseline. In *CVPR*, pages 6536–6545. IEEE Computer Society, 2018.

[LMV+19] Kara Lamb, Garima Malhotra, Athanasios Vlontzos, Edward Wagstaff, Atılım Günes Baydin, Anahita Bhiwandiwalla, Yarin Gal, Alfredo Kalaitzis, Anthony Reina, and Asti Bhatt. Prediction of gnss phase scintillations: A machine learning approach. In *Machine Learning for the Physical Sciences; NeurIPS 2019 workshop*, number arXiv:1910.01570. https://arxiv.org/pdf/1910.01570, 2019.

[LMZW22] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data, 2022.

[LRA20] Guy Leroy, Daniel Rueckert, and Amir Alansary. Communicative reinforcement learning agents for landmark detection in brain images. In Seyed Mostafa Kia, Hassan Mohy-ud Din, Ahmed Abdulkadir, Cher Bass, Mohamad Habes, Jane Maryam Rondina, Chantal Tax, Hongzhi Wang, Thomas Wolfers, Saima

Rathore, and Madhura Ingalhalikar, editors, *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, pages 177–186, Cham, 2020. Springer International Publishing.

[LSM+17] Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6449–6459, 2017.

[LSR+21] Shuangning Li, Matteo Sesia, Yaniv Romano, Emmanuel Candès, and Chiara Sabatti. Searching for consistent associations with a multi-environment knockoff filter. *arXiv preprint arXiv:2106.04118*, 2021.

[LSW+21] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

[LTA+20] Yunzhu Li, Antonio Torralba, Anima Anandkumar, Dieter Fox, and Animesh Garg. Causal discovery in physical systems from videos. *Advances in Neural Information Processing Systems*, 33:9180–9192, 2020.

[LWX+21] Keyu Li, Jian Wang, Yangxin Xu, Hao Qin, Dongsheng Liu, Li Liu, and Max Q. H. Meng. Autonomous navigation of an ultrasound probe towards standard scan planes with deep reinforcement learning, 2021.

[LZWJ16] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016.

[M+15] Volodymyr Mnih et al. Human-level control through deep reinforcement learning. *Nature*, 518:529, Feb 2015.

[MBS19] Fausto Milletari, Vighnesh Birodkar, and Michal Sofka. Straight to the point: reinforcement learning for user guidance in ultrasound, 2019.

[MC00] Subramani Mani and Gregory F Cooper. Causal discovery from medical textual data. In *Proceedings of the AMIA Symposium*, page 542. American Medical Informatics Association, 2000.

[MCL16] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *ICLR*, 2016.

[MLM+19] Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders. *NeurIPS*, 2019.

[MLP21] Scott Mueller, Ang Li, and Judea Pearl. Causes of effects: Learning individual responses from population data, 2021.

[MMG14] Oliver Mattausch, Maxim Makhinya, and Orcun Goksel. Realistic ultrasound simulation of complex surface models using interactive monte-carlo path tracing. 2014.

[MV98] JB Antoine Maintz and Max A Viergever. A survey of medical image registration. *Medical image analysis*, 2(1):1–36, 1998.

[NAS17] NASA. Nasa artemis program. https://www.nasa.gov/artemisprogram, 2017.

[NAS20] NASA. Nasa mars 2020 mission perserverance rover. http://shorturl.at/xzI25, 2020.

[NBS19] Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019.

[NIH21] NIHR. Data sharing guidelines. https://www.nihr.ac.uk/documents/nihr-position-on-the-sharing-of-research-data/12253, 2021.

[NK17] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, (Nips), 2017.

[NK18] Maximilian Nickel and Douwe Kiela. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. *35th International Conference on Machine Learning, ICML 2018*, 2018.

[NLM+18] Alexey A Novikov, Dimitrios Lenis, David Major, Jiří Hladvka, Maria Wimmer, and Katja Bühler. Fully convolutional architectures for multiclass segmentation in chest radiographs. *IEEE transactions on medical imaging*, 37(8):1865–1876, 2018.

[OBG+17] O. Oktay, W. Bai, R. Guerrero, M. Rajchl, A. de Marvao, D. P. O'Regan, S. A. Cook, M. P. Heinrich, B. Glocker, and D. Rueckert. Stratified decision forests for accurate anatomical landmark localization in cardiac images. *IEEE Transactions on Medical Imaging*, 36(1):332–342, Jan 2017.

[OCL+21] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *arXiv preprint arXiv:2111.12525*, 2021.

[OHG+20] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 4 2020.

[OS19] Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

[OVK17] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[PB05] T.E. Page and J.M. Benedetto. Extreme latchup susceptibility in modern commercial-off-the-shelf (cots) monolithic 1m and 4m cmos static random-access memory (sram) devices. In *IEEE Radiation Effects Data Workshop, 2005.*, pages 1–7, 2005.

[PCG20] Nick Pawlowski, Daniel C Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *arXiv preprint arXiv:2006.06485*, 2020.

[Pea99] Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149, 1999.

[Pea09] Judea Pearl. *Causality (2nd edition)*. Cambridge University Press, 2009.

[PSD+20] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[PSWB18] Konstantinos Papangelou, Konstantinos Sechidis, James Weatherall, and Gavin Brown. Toward an understanding of adversarial examples in clinical trials. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–51. Springer, 2018.

[PW21] Sebastian Pölsterl and Christian Wachinger. Estimation of causal effects in the presence of unobserved confounding in the alzheimer's continuum. In Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen, editors, *Information Processing in Medical Imaging*, pages 45–57, Cham, 2021. Springer International Publishing.

[RCP21] Jacob C Reinhold, Aaron Carass, and Jerry L Prince. A structural causal model for mr images of multiple sclerosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 782–792. Springer, 2021.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[RG18] S K Ramakrishnan and K Grauman. Sidekick Policy Learning for Active Visual Exploration. In *EECV*, 2018.

[RHH+10] Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *neuroimage*, 49(2):1545–1558, 2010.

[RLJ20] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications*, 11(1):1–9, 2020.

[Ros58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[RPN12] B. Rahmatullah, A. T. Papageorghiou, and J. A. Noble. Image analysis using machine learning: Anatomical landmarks detection in fetal ultrasound images. In *2012 IEEE 36th Annual Computer Software and Applications Conference*, pages 354–355, July 2012.

[RR83] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[RSdW+18] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. *CoRR*, abs/1803.11485, 2018.

[Rub78] Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34 – 58, 1978.

[Rub05] Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 2005.

[Rub14] Geoffrey D. Rubin. Computed tomography: Revolutionizing the practice of medicine for 40 years. *Radiology*, 273(2S):S45–S74, 2014. PMID: 25340438.

[RVD+22] Hadrien Reynaud, Athanasios Vlontzos, Mischa Dombrowski, Ciarán Lee, Arian Beqiri, Paul Leeson, and Bernhard Kainz. D'artagnan: Counterfactual video generation. *MICCAI*, 2022.

[RVH+21] Hadrien Reynaud, Athanasios Vlontzos, Benjamin Hou, Arian Beqiri, Paul Leeson, and Bernhard Kainz. Ultrasound video transformers for cardiac ejection fraction estimation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 495–505. Springer, Cham, 2021.

[SAM97] Joseph Sill and Yaser S Abu-Mostafa. Monotonicity hints. 1997.

[SBV19] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019.

[SCVH21] Beatriz Garcia Santa Cruz, Carlos Vega, and Frank Hertel. The need of standardised metadata to encode causal relationships: Towards safer data-driven machine learning biological solutions. *Proceedings of CIBB*, page 1, 2021.

[Set12] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[SFMB20] Aishwarya Sivaraman, Golnoosh Farnadi, Todd Millstein, and Guy Van den Broeck. Counterexample-guided learning of monotonic neural networks. *arXiv preprint arXiv:2006.08852*, 2020.

[SG21] Axel Sauer and Andreas Geiger. Counterfactual generative networks. *arXiv:2101.06046*, 2021.

[SGSH00] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[SHK+22] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Di-

ana Mincu, Christina Chen, et al. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*, 2022.

[SHN08] Ramtin Shams, Richard Hartley, and Nassir Navab. Real-time simulation of medical ultrasound from ct images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 734–741. Springer, 2008.

[SJS16] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint arXiv:1606.03976*, 2016.

[SKL+22] Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. *IJCAI*, 2022.

[SLC04] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[SLK18] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

[SLV21] Dídac Surís, Ruoshi Liu, and Carl Vondrick. Learning the predictability of the future. 2021.

[SMS15] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. *ICML*, 2015.

[SN11] Peter Sandercock and Anna. Niewada, Maciej; Czlonkowska. International stroke trial database (version 2). Nov 2011.

[SNDS90] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465 – 472, 1990.

[SRRZ+18] R Sanchez-Romero, J.D. Ramsey, K. Zhang, M. R. K Glymour, B Huang, and C. Glymour. Causal discovery of feedback networks with functional magnetic resonance imaging. *Network Neuroscience*, 2018.

[SRRZG19] Ruben Sanchez-Romero, Joseph D Ramsey, Kun Zhang, and Clark Glymour. Identification of effective connectivity subregions. *arXiv preprint arXiv:1908.03264*, 2019.

[STDE19] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A. Efros. Unsupervised domain adaptation through self-supervision, 2019.

[SWKMM15] Ke Sun, Jun Wang, Alexandros Kalousis, and Stephane Marchand-Maillet. Space-time local embeddings. In *NIPS*. 2015.

[SWTB21] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 519–528, Cham, 2021. Springer International Publishing.

[SWZ+18] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2018.

[TFY20] Long Teng, Zhongliang Fu, and Yu Yao. Interactive translation in echocardiography training system with enhanced cycle-gan. *IEEE Access*, 8:106147–106156, 2020.

[TGS+21] Cristiana Tiago, Andrew Gilbert, Sten Roar Snare, Jurica Sprem, and Kristin McLeod. Generation of 3d cardiovascular ultrasound labeled data via deep learning. *Medical Imaging with Deep Learning-Under Review*, 2021.

[TLYK18a] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing Motion and Content for Video Generation. In *CVPR*, pages 1526–1535. IEEE Computer Society, 2018.

[TLYK18b] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018.

[TP00] Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

[TWT+18] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[TZPG21] Devavrat Tomar, Lin Zhang, Tiziano Portenier, and Orcun Goksel. Content-preserving unpaired translation from simulated to realistic ultrasound images. 3 2021.

[VAK+19] Athanasios Vlontzos, Amir Alansary, Konstantinos Kamnitsas, Daniel Rueckert, and Bernhard Kainz. Multiple landmark detection using multi-agent reinforcement learning. In *International conference on medical image computing and computer-assisted intervention*, pages 262–270. Springer, Cham, 2019.

[VBH+20] Athanasios Vlontzos, Samuel Budd, Benjamin Hou, Daniel Rueckert, and Bernhard Kainz. 3d probabilistic segmentation and volumetry from 2d pro-

jection images. In *The Second International Workshop on Thoracic Image Analysis*, number MICCAI 2020, pages arXiv–preprint, 2020.

[VCB21] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*, 2021.

[VDVVdBDS03] P. Van Dyck, F. M. Vanhoenacker, P. Van den Brande, and A. M. De Schepper. Imaging of pulmonary tuberculosis. *European Radiology*, 2003.

[VKGL21] Athanasios Vlontzos, Bernhard Kainz, and Ciaran M Gilligan-Lee. Estimating the probabilities of causation via deep monotonic twin networks. *arXiv preprint arXiv:2109.01904*, 2021.

[VLT21] Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Re-using adversarial mask discriminators for test-time training under distribution shifts. *arXiv preprint arXiv:2108.11926*, 2021.

[VPT16a] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating Visual Representations from Unlabeled Video. In *CVPR*, pages 98–106. IEEE Computer Society, 2016.

[VPT16b] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating Videos with Scene Dynamics. In *NIPS*, pages 613–621, 2016.

[VRRK20] Athanasios Vlontzos, Henrique Bergallo Rocha, Daniel Rueckert, and Bernhard Kainz. Causal future prediction in a minkowski space-time. *arXiv preprint arXiv:2008.09154*, 2020.

[VRRK22] Athanasios Vlontzos, Hadrien Reynaud, Daniel Rueckert, and Bernhard Kainz. Is more data all you need? a causal exploration. *arXiv*, 2022.

[VSGS21] Athanasios Vlontzos, Gabriel Sutherland, Siddha Ganju, and Frank Soboczenski. Next-gen machine learning supported diagnostic systems for spacecraft.

In *International Workshop on AI for Spacecraft Longevity at IJCAI*. arXiv preprint arXiv:2106.05659, 2021.

[VVdB17] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[VYH+17] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing Motion and Content for Natural Video Sequence Prediction. In *ICLR*, 2017.

[VYZ+17] Ruben Villegas, Jimei Yang, Yuliang Zou, Sungryull Sohn, Xunyu Lin, and Honglak Lee. Learning to generate long-term future via hierarchical prediction. *ICML*, 2017.

[WBBD20] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1160–1169, 2020.

[WBR+20] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive training for improved out-of-distribution detection, 2020.

[WCD21] Rongguang Wang, Pratik Chaudhari, and Christos Davatzikos. Harmonization with flow-based causal inference. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 181–190, Cham, 2021. Springer International Publishing.

[WDZ+19] Erwei Wang, James J. Davis, Ruizhe Zhao, Ho-Cheung Ng, Xinyu Niu, Wayne Luk, Peter Y. K. Cheung, and George A. Constantinides. Deep neural network approximation for custom hardware. *ACM Computing Surveys*, 52(2):1–39, May 2019.

[WPL+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017.

[WTJM20] Sam Witty, Kenta Takatsu, David Jensen, and Vikash Mansinghka. Causal inference using gaussian processes with structured latent confounders. In *International Conference on Machine Learning*, pages 10313–10323. PMLR, 2020.

[YGL+20] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020.

[YGM+19] Xingde Ying, Heng Guo, Kai Ma, Jian Wu, Zhengxin Weng, and Yefeng Zheng. X2ct-gan: Reconstructing ct from biplanar x-rays with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[YHH+21] Xin Yang, Yuhao Huang, Ruobing Huang, Haoran Dou, Rui Li, Jikuan Qian, Xiaoqiong Huang, Wenlong Shi, Chaoyu Chen, Yuanji Zhang, Haixia Wang, Yi Xiong, and Dong Ni. Searching collaborative agents for multi-plane localization in 3d ultrasound. *Medical Image Analysis*, 72:102119, 2021.

[YJVDS18] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[YKBT19] Hui Yang, Soundar Kumara, Satish Bukkapatnam, and Fugee Tsung. The internet of things for smart manufacturing: A review. *IIE Transactions*, pages 1–35, 01 2019.

[YLH+21] Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint arXiv:2106.03721*, 2021.

[YLL+18] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

[YXK+21] Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-congruent training: Towards regression-free model updates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14299–14308, June 2021.

[YXLL21] Haotian Ye, Chuanlong Xie, Yue Liu, and Zhenguo Li. Out-of-distribution generalization analysis via influence function. *arXiv preprint arXiv:2101.08521*, 2021.

[YYY+16] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. *NeurIPS*, 2016.

[ZARX18] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[ZB20] Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcomes. 2020.

[ZBS20] Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020.

[ZDL+21] Anna Zapaishchykova, David Dreizin, Zhaoshuo Li, Jie Ying Wu, Shahrooz Faghihroohi, and Mathias Unberath. An interpretable approach to automated severity scoring in pelvic trauma. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 424–433. Springer, 2021.

[ZDSK+21] Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 279–290, New York, NY, USA, 2021. Association for Computing Machinery.

[ZDT+21] Juntang Zhuang, Nicha Dvornek, Sekhar Tatikonda, Xenophon Papademetris, Pamela Ventola, and James S. Duncan. Multiple-shooting adjoint method for whole-brain dynamic causal modeling. In Aasa Feragen, Stefan Sommer, Julia Schnabel, and Mads Nielsen, editors, *Information Processing in Medical Imaging*, pages 58–70, Cham, 2021. Springer International Publishing.

[Zee64] E. C. Zeeman. Causality implies the Lorentz group. *Journal of Mathematical Physics*, 5(4):490–493, 1964.

[ZGL+22] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. *ArXiv*, abs/2106.06196, 2022.

[ZLG+15] Yefeng Zheng, David Liu, Bogdan Georgescu, Hien Nguyen, and Dorin Comaniciu. 3d deep learning for efficient and robust landmark detection in volumetric data. In *MICCAI 2015*, pages 565–572. Springer, 2015.

[ZWW+20] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020.