Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science
Data Mining Group

Counterfactual Inference on Retina Fundus Images using Deep Structural Causal Models

Master Thesis

Shiqi Liao

Supervisors:
Dr. Vlado Menkovski
Dr. Mitko Veta
Msc. Minartz Koen

EMPTY version

Eindhoven, July 2022

Abstract

The fundus images, a retina photograph, are a potential biomarker mapping to the patient attributes like age, gender, and type-2 diabetes. However, the relationships between the patient attributes and the fundus image appearance are little known and need further explanations. Two limitations in the current method are the confounding between patient attributes and the heterogeneity of the fundus image. The confounding refers to a distorted association(correlation) between two independent variables when an extra variable causes them. For example, older people are likely to have diabetes and vision loss. However, the relationship between diabetes and vision loss does not exist in the population but only in the subpopulation of older people due to the confounder age. The heterogeneity of the fundus image refers to the variation of color, vessel shape, and optic disc location in the fundus image appearance of different individuals.

A counterfactual inference can mitigate these limitations. It is a retrospective hypothesis about an observation under a counterfactual condition. Namely, given an observation, it infers consequence for other features of the observation due to the counterfactual condition and finally results in an alternative status. We can compare the factual observation and the alternative status to deduce the causal relationship between the counterfactual condition to the observation. As the inference is based on the causal relationship and is subject to a specific individual, the counterfactual inference avoids the confounding and heterogeneity.

Recently, a deep structural causal model(DSCM) has been proposed to model the causal relationships and implement the counterfactual inference in silico. The DSCM separates the causal association from the overall association to model the causal relationship, through which the confounding is eliminated. Moreover, the DSCM adopts exogenous noises to keep the implicit individual variation invariant when counterfactual inference. In particular, it adopts normalizing flow and variational inference to infer the exogenous noises for high-dimensional data like images. This enables the DSCM to generate counterfactual inferred images and avoid the heterogeneity of the fundus images.

In this thesis, we construct a custom DSCM to model the causal relationships between patient attributes and fundus images. We validate the custom DSCM on the Maastricht Study. Concretely, we evaluate the image reconstruction and causal inference performance. Furthermore, we implement a sensitivity analysis on the assumed causal relationships. To improve the expressiveness of the custom DSCM on the fundus images, we compare two image preprocessing methods on fundus images and evaluate their impact on model performance. Finally, based on the contrastive counterfactual inference using the custom DSCM on the fundus images, we conclude that aging causes the fundus images to be slightly more yellow, with pixel value increasing in the red and green channels and decreasing in the blue channel.

Preface

This master thesis signifies the end of my master experience of Data Science in engineering at TU/e. I complete my master project within the Data Mining Research Group in the Department of Mathematics and Computer Science and the Medical Image Analysis group (IMAG/e) in the Department of Biomedical Engineering. During these two years, I jumped out of my comfort zone, tried new things, and made new friends. It is scary to confront and accept an unfamiliar challenge. It is suffering when all the effort turns into vain. It is reluctant to admit the gap between myself and the ideal me. However, I gradually learn to handle these emotions, keep the inner peace and keep on working. Day after fay, I know thyself little by little in this repetition. The two year is the start of my exploration into the research and the world. And I believe it is not an end to it.

My master project is under the supervison of Dr. Vlado Menkovski, Dr. Mitko Veta and Msc. Minartz Koen. Dr. Mitko Veta drafted the idea and had a weekly meeting with me to discuss my doubts and analyze the experiment results. Dr. Vlado Menkovski invited me to his research group and always encouraged me to communicate with my peers. And Msc. Minartz Koen helped me to finetune the model and revised my thesis. I want to thank them for their instruction and accompanying along the way.

Furthermore, I want to thank my friends. They tried to understand me and listen to me no matter the distance. I have learned a lot from them and I am so lucky to have them in my life.

Lastly, I want to thank my parents for always supporting me and accompanying me no matte my choice.

Contents

List of Figures

List of Tables

Listings

Chapter 1

Introduction

The retina is the only tissue in the body that allows for non-invasive simultaneous visualization of neuronal and vascular tissue from brains[3]. As a result, it does not only function as a tissue layer of human eyes, but also functions as a part of the brain's vascular system and hence can reflect aging and systemic disease[36]. Many photographs have been introduced to clinics to observe and monitor the retina. The most common and cheapest retinal photograph is color fundus images that record the appearance of a patient's retina in 2D imaging modality[45].

A valuable application of retina in ophthalmology, eye science, is an efficient biomarker mapping to systemic indices of healthy aging and disease[7, 27, 55, 57, 33, 29]. Recently, the retina has shown its potential relation with patient attributes using deep neural networks. With color fundus image, deep neural networks have successfully predicted demographic information such as age and gender[38], and systemic diseases like chronic kidney disease[59] and type-2 diabetes[17]. This finding indicates the potential of the retina as a biomarker for patient attributes.

To clarify the relationships between patient attributes and retinal structure, traditionally, we derive the relationships between retinal morphology and patient attributes using statistical modeling, such as multivariable regression[12, 58]. As medical data availability has increased, generative models can derive previously hidden patterns in large volumes of data by interpreting differences in generated data under different generation conditions[40, 31]. Recently, a pioneering development within the generative model is the introduction of causality[48], one example of which is the deep structural causal model[37]. It allows for comparing the factual and counterfactual inferred images, a retrospective hypothesis about an observation under a counterfactual condition. The difference could be interpreted as the causal explanation of the impact of the condition.

1.1 Motivation

In clinical medicine, transparency in decision-making is of vital importance[1, 23]. Without solid evidence and explanation in the decision-making process, the patients may suffer unnecessary healthcare and even receive wrong treatments[41]. Hence despite deep neural network image classifiers having high accuracy, the connection between the image and class cannot be trusted and explained due to their unknown decision-making process and potential biases. The classifiers of gender, age, and Type-2 diabetes on fundus images suffer the same problem[38, 17]. A novel attempt by Philipp[31] is to generate realistic fundus images conditioned on patient attributes and investigate the relationships between patient attributes and generated image appearance. The generative model can generate the fundus images based on the statistical pattern found in the data and the statistical pattern is easier to be observed when it presents itself in the generated images.

However, two challenges exist for the current method. One challenge is confounding between patient attributes. Uncontrolled confounders would distort the association between a patient attribute and the generated images, so the relationship between the patient attribute and the

generated image appearance is disturbed. Another challenge is that the fundus images are highly heterogeneous with a wide range of appearances. Namely, fundus images of different people are diverse in color, vessel shape, and optic disc location, etc. So contrastive generated fundus images cannot explicitly reflect the relationships between patient attributes and fundus images.

In order to mitigate these limitations, we expect to construct a causal model that model causal relationships rather than associational relationships between patient attributes and fundus images. The model should control confounding between patient attributes and fundus images. Next, we want to use the causal model to generate a counterfactual inferred fundus image based on a factual fundus image of an observational participant. The counterfactual inferred fundus image should be what the fundus image would be like if the participant had a counterfactual patient attribute. Finally, by comparing the factual fundus image and the counterfactual inferred image, we want to explore the causal relationships between the patient attributes and fundus image appearance.

1.2 Contributions

Our project focuses on contrastive counterfactual inference on fundus images. We apply our experiments on the Maastricht study[1], an observational prospective population-based cohort study. Furthermore, our main contributions are:

- We develop a custom deep structural causal model(DSCM) on fundus images. It models the causal relationships between patient attributes. Furthermore, it generates realistic fundus images by counterfactual inference. With contrastive counterfactual inferred fundus images, we deduce the causal relationships between age and fundus images. We also implement a sensitivity analysis on the custom DSCM design.
- We compare different image preprocessing methods on fundus images and their impact on DSCM's image reconstruction and counterfactual inference performance. We mainly focus on two types of image preprocessing methods, contrast normalization[13] and vessel segmentation[43].

1.3 Outline

The remainder of this work is presented as follows. Chapter 2 contains the problem statement. In Chapter 3, the preliminary is given. In chapter 4, the related work of causal models supporting counterfactual inference on images is summarized. In chapter 5, the Dataset and Methods are described. In Chapter 6, we present the experiment results. Finally, in Chapter 7, we give our conclusions, limitations, and possible future work.

[1] https://www.demaastrichtstudie.nl/

Chapter 2

Problem statement

In this section, we explain the underlying tasks of counterfactual inference on fundus images. One task is to eliminate confounding with Causal Bayesian Network. The other task is to make an individual inference with DSCM to avoid heterogeneity issue. Furthermore, we elaborate on the main research question of this project.

2.1 Task 1: Causal Bayesian Network Design to control confounding

Confounding is commonly referred to as a "mixing of effects"[54] and occurs when the effects of the factor under study on a given outcome are mixed in with the effects of another factor, resulting in a distortion of the true relationship. The confounding can only be controlled with limitation to a subpopulation if we model statistical dependence relationships. However, we can control the confounding on population level if we model causal relationships. Causal Bayesian Network by Pearl[6] can model the causal relationships between variables in a directed acyclic graph. By deleting the directed edges representing the causation from cause to effect, we can block the causal relationships that result in confounding.

We aim to design a Causal Bayesian Network that models the causal relationships between patient attributes and fundus images. Furthermore, we can eliminate confounding by intervening on the model.

2.2 Task 2: Counterfactual Inference to avoid heterogeneity of fundus images

Fundus images are heterogeneous in color, vessel shape, optic disc location, and due to observation bias in photographing, etc. See Fig.2.1 for an example of heterogeneous fundus images and Fig.2.2 for a fundus image structure. When analyzing the difference in fundus images of participants with different attributes, it is difficult to decide if the effect of the attributes or heterogeneity is attributed to the difference. Ideally, we can compare fundus images of a participant without and with type-2 diabetes induced by a high-fat diet to deduce the relationship between type-2 diabetes and fundus images. However, such an unethical human experiment is forbidden.

We can implement such an unethical human experiment in silico with a deep structural causal model(DSCM) by Pawlowski[37]. The DSCM uses deep learning to quantitatively model causal relationships, and users can modify the causal relationships arbitrarily. So we can use the DSCM to model the outcome under the different cause statuses. It also includes an implicit individual variation to preserve essential individual information. The implicit individual variation enables the DSCM to infer the outcome for a specific individual. And we name such inference as counterfactual inference.

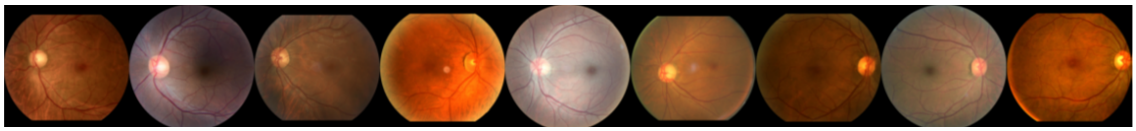


Figure 2.1: Heterogeneous fundus image, image from [31]

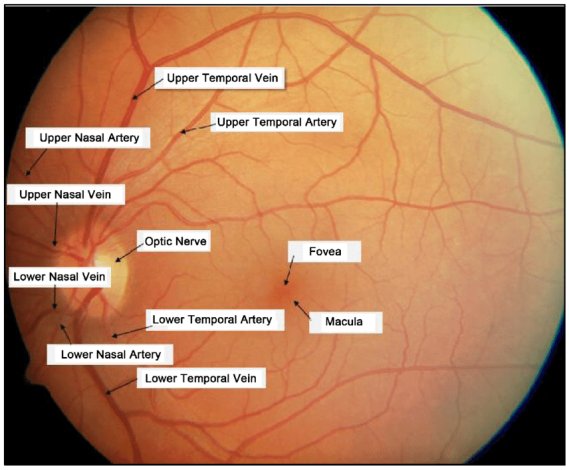


Figure 2.2: Fundus structure

We aim to construct a custom deep structural causal model on fundus images and generate counterfactual inferred fundus images.

2.3 Formulation

Given Maastricht study dataset $\mathbf{X} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\right\}$, which records variables, age(a), gender(g), type-2 diabetes status(d), and fundus image(i) for each participant $\boldsymbol{x}^{(n)}$, $\boldsymbol{x}^{(n)} = \left\{a^{(n)}, g^{(n)}, d^{(n)}, i^{(n)}\right\}$.

Firstly, we design a custom Causal Bayesian Network (G, p) to model the causal relationships between variables. G denotes a directed acyclic graph(DAG), which uses V, nodes, to denote the set of variables $V = \{a, g, d, i\}$ and use E, edges, to denote the set of causation from a cause to an effect. And p indicates joint probability distribution $p\left(V_1, \ldots, V_k\right) = \prod_{V_i \in V} p\left(V_i \mid \mathbf{Pa}(V_i)\right)$, where $\mathbf{Pa}(V_i)$ indicates the parent nodes of node V_i in G.

Secondly, we design a custom structural causal model(DSCM) $\mathfrak{G} = (\mathbf{S}, \mathbf{p}(\boldsymbol{\epsilon}))$ to quantitatively model the causal relationships. $\mathbf{S}$ denotes a set of quantitative causal structural equations $V_i := f_i(\epsilon_i; \mathbf{Pa}(V_i)), V_i \in V$. And the equation should model the causation between the causes $\mathbf{Pa}(V_i)$ and the effect V_i. Also, $\mathbf{p}(\boldsymbol{\epsilon})$ denotes a joint distribution over exogenous noises. These noises represent implicit individual variations.

Finally, we use the DSCM to generate fundus images under different counterfactual conditions for an individual participant v. We first infer the implicit individual variations $p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})$. Then according to the desired counterfactual conditions, we intervene on the DSCM to get a modified DSCM $\widetilde{\mathfrak{G}} = \left(\widetilde{\mathbf{S}}, p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})\right)$. After that, we generate the counterfactual inferred fundus images by recomputing the modified DSCM $\widetilde{\mathfrak{G}}$.

2.4 Research question

We construct a causal model that quantitatively models the causal relationships between patient attributes and fundus images. And we use the model to generate a realistic fundus image for a specific participant with a counterfactual patient attribute. By comparing the counterfactual inferred and factual fundus image, we expect to answer the following question:

What is the causal relationships between the patient attributes and the fundus image appearance?

Chapter 3

Preliminaries

We first present the research background, including type-2 diabetes, fundus images. Next, we elaborate on the field of causality, particularly the three ladders of causal hierarchy and their corresponding causal models. We specifically emphasize the structural causal model(SCM) that is capable of counterfactual inference. As an example, we make inferences on different levels of causal hierarchy on a toy linear structural causal model. After that, we introduce deep learning, which can be embedded into an SCM as automatically learned structural equations. Since we are interested in generating a fundus image, high-dimension data, we elaborate on the generative models under unsupervised deep learning. We introduce Normalizing Flow and Variational Autoencoder in detail as they are adopted in our experiment. We also briefly introduce other representative models, Generative adversarial network and diffusion, as a base for summary of causal models supporting counterfactual inference on images in section 4.

3.1 Background on Type-2 Diabetes and Fundus images

Diabetes is a metabolic disease usually diagnosed with high blood sugar and has two main types: type-1 diabetes and type-2 diabetes[35]. About 90-95% of diabetes patients have type-2 diabetes[19]. These patients don't produce enough insulin or can't use it well. So they stay in the status lacking insulin and high blood sugar as a result. Over time, they will suffer serious health problems like heart disease, vision loss, and kidney disease. Furthermore, according to population health surveys in Canada and Europe, older people and males are more likely to have type-2 diabetes[14, 4].

Fundus images are a type of retinal photograph. It records the appearance of a patient's retina in a color 2D imaging modality[27]. They are usually used by eye doctors to monitor the progression of specific eye diseases. Besides, they are also used to record abnormalities of systemic disease process affecting eyes, such as diabetes, age-macular degeneration(AMD), glaucoma, and multiple sclerosis. In particular, for type-2 diabetes patients, regular fundus screening examinations are necessary because vision loss due to diabetes can be prevented by retinal laser treatment if vision loss is diagnosed early[28]. The main structures visualized on a fundus image are an optic disc, a macula, temporal and nasal vessels coming from a brain's vascular system[36].

3.2 Causal Hierarchy and Causal model

It is a commonplace in a scientific discussion that correlation does not imply causation[42]. Correlation is defined as two variables that appear to be so closely associated that one is dependent on the other. However, causation explicitly applies to cases where one variable causes another one. There are some amusing examples of the dangers of inferring causation from correlation. For example, Nobel laureates correlate with chocolate consumption[30] as shown in Fig 3.1. Despite

this strong correlation, it is wrong to conclude that the high chocolate consumption in Switzerland has somehow caused the high Nobel laureates winners in Switzerland, nor that the high Nobel laureates winners in Switzerland have somehow caused the high chocolate consumption in Switzerland.

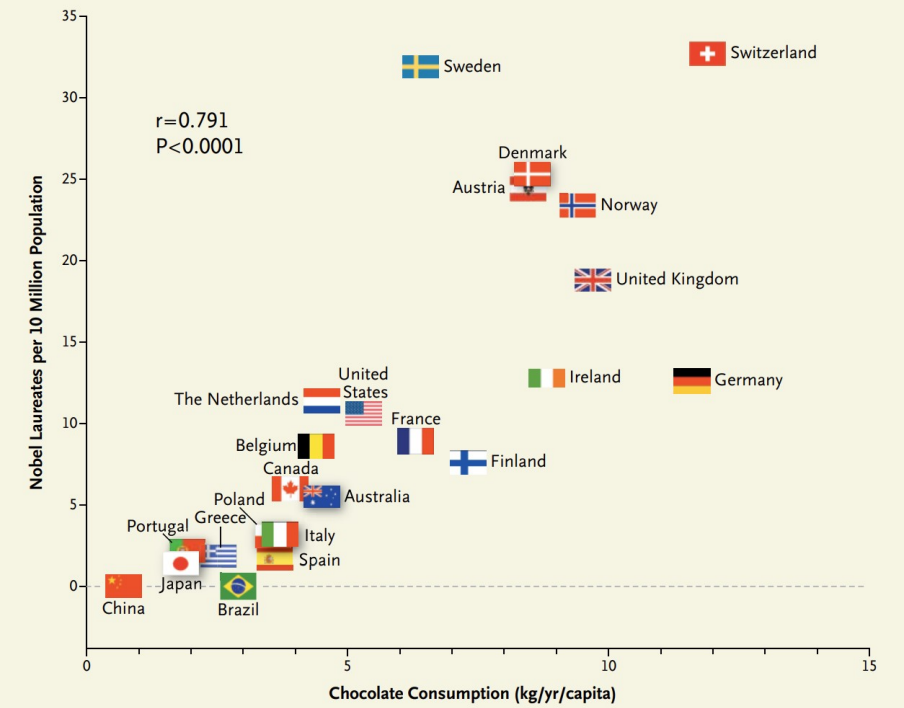


Figure 3.1: Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population. Image from [30]

To understand the correlation between Nobel laureates and high chocolate consumption when there is no causal relation between them, we have to explore the underlying causal relationship. The correlation is observed when there is an unobserved confounding, GDP, on them. A country with a higher GDP has richer citizens who are affordable of chocolate and encourage scientific research. So GDP has positive effects on both chocolate consumption and Nobel laureates. As a result, it falsely demonstrates an apparent association between chocolate consumption and Nobel laureates when there is no causal relation between them.

The introduction of causality can deepen our understanding of the relationship between factors, and thus such confounding could be eliminated. To formulate the causation and furthermore to license the causal reasoning and causal inference, Judea Pearl[39] comes up with a causal hierarchy with three ladders: 1) association 2) intervention 3) counterfactuals. We present the Causal Hierarchy in Table 3.1. Each level presents a understanding of the relationship. It comes with a characteristic type of question named causal reasoning. The process of answering the causal reasoning is named causal inference. Furthermore, a higher ladder has a deeper understanding and thus has more complex causal reasoning. A representative model from some level can not only model the relationship and answer the causal reasoning on the corresponding level but also lower levels. In the following sections, we will clarify each level together with a representative causal model. Finally, we make associational, interventional, and counterfactual inferences on a toy linear structural causal model.

Table 3.1: Causal hierarchy

Level	Typical Activity	Typical Question	Examples
1. Association $P(Y\|X)$	Seeing	How would seeing X change my belief in Y?	Are people over 60 years old more likely to have type-2 diabetes?
2. Intervention $P(Y\|do(x))$	Doing	What happens if I do?	What happens to fundus images if all participants have type-2 diabetes?
3. Counterfactuals $P(Y_x\|x', y')$	Retrospection	What would happen if I had done differently?	What would happen to my fundus images if I had been diagnosed with type-2 diabetes?

3.2.1 Association and Bayesian Network

We call the first level Association because it describes only statistical dependence relationships, which can be observed from the raw observational data. An example of associational reasoning on this level is: does observing an old clinic visitor make him more likely to have diabetes? Such reasoning can be answered directly from the observational data using conditional probability $p(Y|X)$. The answer is named associational inference. As there is no modification on the observational data, the typical activity of this level is seeing what it is. Questions at this layer are at the bottom level of the hierarchy as causal information is unnecessary.

We usually use Bayesian Network to represent the association level. As a directed acyclic graph(DAG) is the base of the Bayesian Network, we first give the definition of DAG and additional Local Markov Assumption and Minimality Assumption.

Definition 3.2.1 (Directed acyclic graph(DAG)) *A directed acyclic graph(DAG) is a directed graph:* $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ *with the following graph terminology.*

- $\mathbf{V}$ *is a set of nodes*
- $\mathbf{E}$ *is a set of directed edges between the nodes*
- *no cycle should be included in a DAG*
- $\mathbf{Pa}(\mathbf{V})$ *denotes Parents of a node* $\mathbf{V}$*, also named nodes that have an edge pointing to* $\mathbf{V}$
- $\mathbf{Ch}(\mathbf{V})$ *denotes Children of a node* $\mathbf{V}$*, also named nodes that have an edge pointing from* $\mathbf{V}$
- *A path between node* i *and node* j *is a sequence of distinct nodes* $(\mathbf{i}, \ldots, \mathbf{j})$ *such that each two consecutive nodes are adjacent*
- *A directed path between node* i *and node* j *is a path where all edges point towards j, i.e.* $\mathbf{i} \rightarrow \ldots \rightarrow \mathbf{j}$*. we call a path that is not a directed path a non-directed path.*

Assumption 3.2.1 (Local Markov Assumption) *Every node V in a DAG is conditionally independent of nodes that are not* $\mathbf{Ch}(\mathbf{V})$*, given* $\mathbf{Pa}(\mathbf{V})$*.*

Assumption 3.2.2 (Minimality Assumption) *Adjacent nodes in the DAG are dependent.*

Given a DAG under such two assumptions, we can apply the chain rule to factorize a joint distribution of the variables. Such a DAG and factorization construct a Bayesian network as Definition 3.2.2.

Definition 3.2.2 (Bayesian network) *Given a set of random variables* $(V_1, ..., V_k)$ *with joint distribution* $p(V_1, ..., V_k)$*. A DAG* G *can represent each random variable* V_i *as node* i *and represent their dependence as edges. The observational joint distribution* $p(V_1, ..., V_k)$ *can factorize over* G *as* $p(V_1, \ldots, V_k) = \prod_{V_i \in V} p(V_i \mid \mathbf{Pa}(V_i))$*. A Bayesian network is the tuple* (G, p)*.*

In Bayesian Network, the Local Markov Assumption refines the dependencies by making a node only dependent on its child nodes given its parent nodes, Minimality Assumption endows edges with dependence between variables. And we can compute conditional probability based on the dependence. Take a toy Bayesian Network in Fig 3.2 to compute the conditional probability $p(V_3|V_1)$ as an example. The V_1 and V_3 are confounded by V_2. We can factorize locally along all paths between nodes V_1 and V_3 to get an observational joint distribution of all variables over the paths, $p(V_1, V_2, V_3) = p(V_2)p(V_1|V_2)p(V_3|V_1, V_2)$. Then we marginalize out the variables between node X and Y from conditional joint distribution as follows:

$$p(V_3 \mid V_1) = \sum_{V_2} \frac{p(V_1, V_2, V_3)}{p(V_1)}, \tag{3.1}$$

where we use sum as we assume the variables are discrete and we should use integral for continuous variables.

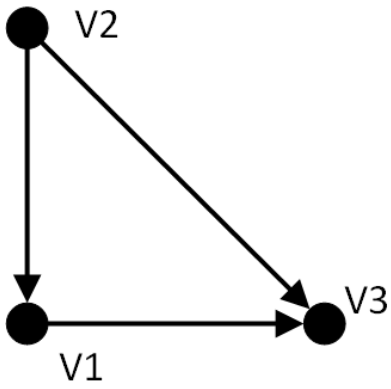


Figure 3.2: Toy Bayesian Network

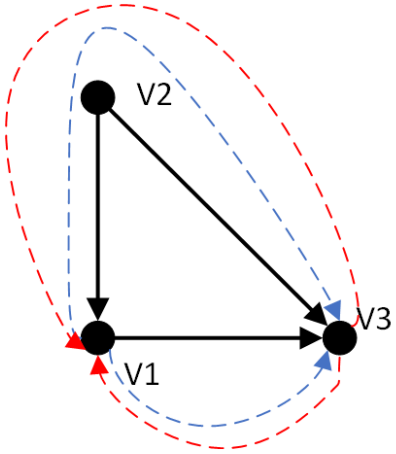


Figure 3.3: Toy Bayesian Network with symmetric association flows

Note that as dependence is mutual, the edges in the toy Bayesian Network can be flipped and thus the observational joint distribution can also be factorized symmetrically as follows:

$$p(V_1, V_2, V_3) = p(V_1)p(V_2|V_1)p(V_3|V_1, V_2) = p(V_3)p(V_2|V_3)p(V_1|V_2, V_3) \tag{3.2}$$

Such factorization follows paths ($V_1 \to V_3$, $V_1 \to V_2 \to V_3$) in blue flows or paths ($V_3 \to V_1$, $V_3 \to V_2 \to V_1$) in red flows as shown in Fig 3.3. We can intuitively treat such paths from node V_1 to V_3 as association flows, along which the dependence between node V_1 to V_3 is determined by factorization. Obviously, the association flows are symmetric.

Finally, we can answer the associational reasoning in Bayesian Network using conditional probability.

3.2.2 Intervention and Causal Bayesian Network

The second level, Intervention, has a higher rank than the Association because it involves modification to the observational data rather than just observing it. Namely, it involves a do operation to modify the DAG to utilize the observational data to infer on the population level. A typical

interventional reasoning is: What happens to fundus images in the observed population if all participants have type-2 diabetes? Such questions cannot be answered from observational data alone on association level, because the relationship between type-2 diabetes and fundus images involves a confounder, age. Under different age ranges, the type-2 diabetes status may differ substantially. Hence the relationship between type-2 diabetes and fundus images changes accordingly. On the intervention level, we can use the interventional distirbution $p(Y \mid do(x))$, which denotes the probability of event $Y = y$ given that we intervene on X as x to answer the interventional reasoning. This is also named interventional inference.

Within the following content, we will first define cause and add an extra assumption to DAG so that it introduces causation into Bayesian Network. Then we explain how we separate the overall association into causal association and non-causal association in the view of causation. The non-causal association is confounding in our case. Next, we elaborate on the idea of blocking the non-causal association so that the overall association is equal to the causal association. After that, we define a causal Bayesian Network where a do operation can block the non-causal association on the population level and compute interventional distribution $p(Y \mid do(x))$ as the causal association on the population level. Lastly, we use the toy Bayesian Network in Fig 3.2 as an example to explain the difference between the causal association and overall association.

Definition 3.2.3 (Cause) *A variable X is said to be a cause of a variable Y if Y can change in response to changes in X. And the variable Y is named an effect.*

Assumption 3.2.3 (Causal Edges Assumption) *In a DAG, every parent is a direct cause of all its children.*

A DAG under Causal Edges Assumption is named causal DAG, where the directed edges take the meaning of causation. Furthermore, the directed edges are asymmetrical. The causation can flow along the directed paths from a cause to a further effect. Different from the directed paths, other paths allow the association to flow along but not causation. Note that though the directed edges represent causation, the dependence between the adjacent nodes still holds. Now we can analyze association from a causal view. The overall association between two variables can be divided into a causal and non-causal association. To explain this claim, we look into the toy Bayesian Network in Fig 3.2 again with the Causal Edges Assumption. The association flows along two paths($V_1 \to V_3$, $V_1 \leftarrow V_2 \to V_3$) from node V_1 to V_3. The first path $V_1 \to V_3$ is a directed path and the causal association flows this path from V_1 to V_3. The second path $V_1 \leftarrow V_2 \to V_3$ is not a directed path and thus the non-causal association flows this path from V_1 to V_3. From this claim, we can say that causation is a sub-category of association.

Given that we can measure the overall association via factorization, how can we measure the causal association as a part of the overall association? We can achieve this by guaranteeing only a causal association between a cause and an effect by blocking the non-causal association. As the non-causal association is the dependence flows along the non-directed paths, a simple solution is to condition on some nodes in the non-directed paths to make the cause and effect conditionally independent with respect to the non-directed paths. For example, in a toy Bayesian network in Fig 3.2, we can condition on V_2 to make V_1 and V_3 conditionally independent with respect to the path $V_1 \leftarrow V_2 \to V_3$. Such conditional independence can be proved as Eq 3.3.

$$p(V_3, V_1|V_2) = \frac{p(V_1, V_2, V_3)}{p(V_2)} = \frac{p(V_2)p(V_1|V_2)p(V_3|V_2)}{p(V_2)} = p(V_1|V_2)p(V_3|V_2) \tag{3.3}$$

With condition on V_2, the causal association between node V_1 and V_3 can be represented with conditional probability $p(V_3|V_1, V_2 = v_2)$ as in Eq 3.4. However, this causal association is restricted to a subpopulation where $V_2 = v_2$ rather than the whole population.

$$p(V_3|V_1, V_2 = v_2) = \frac{p(V_1, V_3|V_2 = v_2)}{p(V_1|V_2 = v_2)} \tag{3.4}$$

To compute causal association on the population level, we have to jump out of the association level and take the tool of causation. Now that the directed edges in causal DAG denote the asymmetric causation towards an effect node, we can directly isolate a cause from the non-causation flow by deleting the incoming edges towards the cause. Then we factorized over the post-interventional causal DAG to compute the causal association between the cause and the effect. The intervention is not restricted to some subpopulation so the computed causal association is on the populaiton level. To achieve this, The causation along other causal edges should remain the same before and after the intervention. This requirement is satisfied by Modularity Assumption, and we define it as follows:

Assumption 3.2.4 (Modularity) *Given the intervention on a set of nodes $V_{\mathbf{W}}$, setting them to constants. Then we have the following modifications to the probability distribution.*

- *For all intervened nodes $V_i \in V_{\mathbf{W}}$, set $p(v_i \mid \mathbf{Pa}(V_i))$ to 1 if v_i is equal to the set value by intervention and set $p(v_i \mid \mathbf{Pa}(V_i))$ to 0 if not.*
- *The unintervened nodes $V_i \notin V_{\mathbf{W}}$ keep the same $p(V_i \mid \mathbf{Pa}(V_i))$.*

A Bayesian Network under Causal Edges Assumption and Modularity Assumption is named a causal Bayesian Network. We define a causal Bayesian Network as follows:

Definition 3.2.4 (Causal Bayesian Network) *Given a Bayesian Network (G, p) Where G is a causal DAG $G(V, E)$, (G, p) is a causal Bayesian network where for any $V_i \in \mathbf{V}$*

$$p(V_1, \ldots, V_k \mid \mathrm{do}(V_{\mathbf{W}} = v_{\mathbf{W}})) = \begin{cases} \prod_{V_i \in \mathbf{V} \setminus V_{\mathbf{W}}} p(V_i \mid \mathbf{Pa}(V_i)) & \text{if } V_{\mathbf{W}} = v_{\mathbf{W}} \\ 0 & \text{otherwise} \end{cases}, \tag{3.5}$$

where $\mathrm{do}(V_{\mathbf{W}} = v_{\mathbf{W}})$ denotes the intervention, aka. do-operator and $p(V_1, \ldots, V_k \mid \mathrm{do}(V_{\mathbf{W}} = v_{\mathbf{W}}))$ denotes the corresponding interventional distribution. The interventional distribution is the factorization over the intervened DAG after the do-operator.

To give an intuition on how the do-operator blocks the non-causal association, we again take the toy Bayesian Network as an example. This time we define it as a causal Bayesian Network in Fig 3.4 and perform a do-operator $do(V_1 = v_1)$.

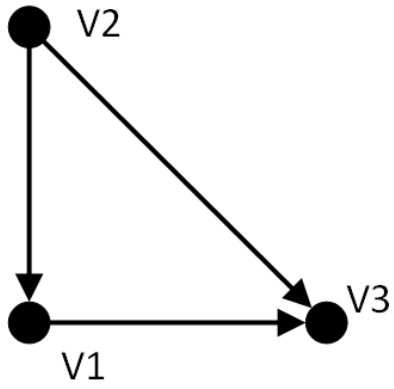


Figure 3.4: Toy causal Bayesian Network

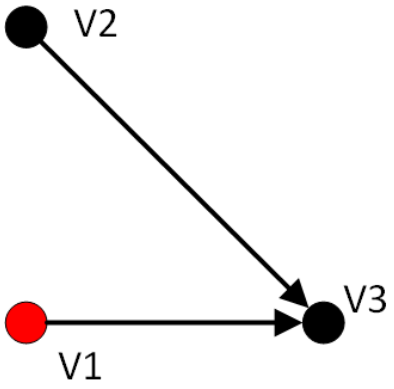


Figure 3.5: Toy causal Bayesian Network after intervention

Initially, the observational joint distribution with respect to the causal Bayesian Network is as follows:

$$p(V_1, V_2, V_3) = p(V_2)p(V_1 \mid V_3)p(V_3 \mid V_1, V_2) \tag{3.6}$$

Then we perform an intervention $do(V_1 = v_1)$. Following the modification to the factorization in Eq 3.5, we deduce the interventional distribution $p(V_2, V_3|do(V_1 = v_1))$ in Eq 3.7.

$$p(V_2, V_3|do(V_1 = v_1)) = p(V_2)p(V_3|V_2, V_1 = v_1) \tag{3.7}$$

Next, we marginalize out V_2 to get $p(V_3|do(V_1 = v_1))$ in Eq 3.8.

$$p(V_3 \mid do(V_1 = v_1)) = \sum_{V_2} p(V_3 \mid V_2, V_1 = v_1)p(V_2) \tag{3.8}$$

The intervened causal Bayesian Network after this intervention is shown in Fig 3.5. Intuitively, we can see that the non-causal association flow along the path $V_1 \leftarrow V_2 \rightarrow V_3$ is blocked. Hence the node V_1 and the node V_3 are independent with respect to the path $V_1 \leftarrow V_2 \rightarrow V_3$. By factorizing over the intervened causal Bayesian Network, we get the new overall association equal to the pre-interventional overall association minus the non-causal association. This new overall association is the causal association as a result.

Next, we explain the difference between the causal association $p(V_3|do(V_1 = v_1))$ and the overall association $p(V_3|V_1 = v_1)$. We deduce the overall association as in Eq 3.9.

$$p(V_3 \mid V_1 = v_1) = \sum_{V_2} p(V_2, V_3 \mid V_1 = v_1) = \sum_{V_2} p(V_3 \mid V_2, V_1 = v_1)p(V_2 \mid V_1 = v_1) \tag{3.9}$$

Comparing Eq 3.9 and Eq 3.8, we can conclude the difference between causal association $p(V_3|do(V_1 = v_1))$ and the overall association $p(V_3|V_1 = v_1)$ is the difference between $p(V_2)$ and $p(V_2 \mid V_1 = v_1)$. This implies the confounding in the overall association is attributed to the dependence between the node V_2 and the node V_1. By replacing $p(V_2 \mid V_1 = v_1)$ with $p(V_2)$ in the marginalization, the causal association avoids confounding.

Finally, we can answer the interventional reasoning by computing the causal association on the population level with interventional distribution $p(Y \mid do(x))$. Note that we only mention the confounding case in this chapter as it is what our project focuses on. And there are other non-causal associations flowing along different types of paths to be blocked. Pearl defines these as blocked paths and calls a cause and an effect whose non-causal association flows are all blocked as d-separated. The causal Bayesian Networks can achieve d-separation by methods like frontdoor adjustment and backdoor adjustment. Please refer to [32] for more information.

3.2.3 Counterfactuals and Structural Causal Model

Finally, the top level is called Counterfactuals. In contrast to intervention, where inference is based on population distribution, counterfactual expects to infer on the individual level with updated information deduced from the observational evidence. A typical counterfactual reasoning is "What would happen to my fundus images if I had been diagnosed with type-2 diabetes?". To answer the question, we have an expression $p\left(Y_x \mid x', y'\right)$ which stands for the probability of event $Y = y$ had X been x, given that we have observed X used to be x' and and Y used to be y'. Such expression is named counterfactual inference and can be computed by Structural Causal Model.

As we mentioned in section 3.2.2, a causal Bayesian Network can encode the statistical dependence and causal assumption well. And it can represent the causal association by the do-operator. However, it cannot include implicit individual variation when inferring on the individual level, so it can not make the counterfactual inference. The implicit individual variation can explain the heterogeneity as different individuals have different implicit individual variations. Namely, it can explain the unchanged part when inferring the same individual under different counterfactual conditions. Pearl comes up with a Structural Causal Model(SCM) to achieve the counterfactual inference. It supports counterfactual inference by clarifying the causal association quantitatively and adopting exogenous noises as the implicit individual variations. Within the following content,

we define SCM in definition 3.2.5 and three steps in counterfactual inference in definition 3.2.6. Then we explain how it satisfies the Causal Edges Assumption and Modularity Assumption and utilize exogenous noises to make inferences on the individual level.

Definition 3.2.5 (Structural causal model(SCM)) *Given a causal Bayesian network (G, p) where G is a causal DAG $G(V, E)$, each variable V_i for $V_i \in V$ is assigned to a structural equation of its parents $\mathbf{Pa}(V_i)$ and a noise term ϵ_i as follows:*

$$V_i := f_i\left(\epsilon_i; \mathbf{Pa}(V_i)\right) \tag{3.10}$$

Define a collection $\mathbf{S} = (f_1, \ldots, f_K)$ of structural equations and a joint distribution $\mathbf{p}(\epsilon) = \prod_{k=1}^{K} p\left(\epsilon_k\right)$ over mutually independent exogenous noise variables $\forall i \neq j : \epsilon_i \perp\!\!\!\perp \epsilon_j$.

Then $(\mathbf{S}, \mathbf{p}(\boldsymbol{\epsilon}))$ is a structural causal model.

In a structural causal model, we define a function f_i mapping from an exogenous noise ϵ_i and parent nodes $\mathbf{Pa}(V_i)$ to node V_i. We use the exogenous noise ϵ_i to represent the implicit individual variation. It is named an exogenous noise for it is external to the causal model and is added artificially to explain the heterogeneity. Based on such intention, it has no parent node. Also note that $:=$ has causal information rather than associational information, and thus the mapping is asymmetric. In this way, the structural causal model satisfies the Causal Edges Assumption.

The counterfactual inference follows the three steps below.

Definition 3.2.6 (Three steps in Counterfactual inference) *Counterfactual inference consists of three necessary steps:*

- *Abduction: Predict the 'implicit individual variation' (the exogenous noise, ϵ) that is compatible with the individual observation,* v*, i.e. infer $p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})$.*
- *Action: Perform the desired intervention (e.g. $\mathrm{do}\left(v_k := \widetilde{v}_k\right)$) to the original structural equations $\mathbf{S}$ and result in a modified structural equations $\widetilde{\mathbf{S}}$. The modified structural equations $\widetilde{\mathbf{S}}$ and inferred exogenous noise $p_{\mathfrak{G}(\epsilon|\mathrm{v})}$ form a modified SCM $\widetilde{\mathfrak{G}} = \mathfrak{G}_{\mathbf{v};\mathrm{do}(\widetilde{v}_k)} = \left(\widetilde{\mathbf{S}}, p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})\right)$.*
- *Prediction: Compute the value of interested variables $v_i \in \mathrm{v}$ based on the modified SCM $\widetilde{\mathfrak{G}}$.*

For the abduction step, even though it is stated in definition 3.2.5 that function f is asymmetric. We can infer the value of the exogenous noise ϵ because mutual dependence still holds in causation. And for the action step, the intervention $\left(\mathrm{do}\left(v_i := \widetilde{v}_i\right)\right)$ replaces the structural equation $v_i := f_i\left(\epsilon_i; \mathbf{Pa}(v_i)\right)$ with $v_i := \widetilde{v}_i$. In this way, $p\left(v_i = \widetilde{v}_i \mid \mathbf{Pa}(v_i)\right) = 1$ and $p\left(v_i \neq \widetilde{v}_i \mid \mathbf{Pa}(v_i)\right) = 0$. Furthermore, as the structural equations and exogenous noises are independent respectively, the intervention on one variable will not change other structural equations. So the intervention satisfies the Modularity Assumption, and it can isolate the causal association from the intervened variable toward its effects. For the action step, the modified SCM $\widetilde{\mathfrak{G}}$ consists of the intervened structural equations $\widetilde{\mathbf{S}}$ and inferred exogenous noises $p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})$. The intervened structural equations $\widetilde{\mathbf{S}}$ models the causation flow and the inclusion of the inferred exogenous noises $p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})$ can preserve heterogeneous information for an individual observation. Given such a modified SCM $\widetilde{\mathfrak{G}}$, we can model the causation from the intervened variable to its effects by recomputing the variables. As a result, it achieves counterfactual inference on the individual level and answers the counterfactual reasoning $p\left(Y_x \mid x', y'\right)$.

Note that the counterfactual level is the highest causal hierarchy, so it can explain the relationships on lower levels, association and intervention. As a representative model for the counterfactual level, the SCM can also model the overall association and causal association on the population level. We can assume exogenous noises follow a prior distribution and abandon the abduction step to make the exogenous noises work as an implicit population variation rather than an implicit individual variation. Under this assumption, users can model the overall association by computing on the original SCM $\mathfrak{G}$ with exogenous noises sampling from the prior distribution. Furthermore, users can model the causal association by implementing the action and prediction step with exogenous noises sampling from the prior distribution.

3.2.4 Causal Inference on an Example Model

In this section, we perform causal inference on an example linear SCM. Firstly, we will show how confounding influences the association. Next, we will implement intervention to eliminate confounding. Then we follow the three steps in definition 3.2.6 to implement counterfactual inference.

Firstly, we construct an example linear structural causal model in Fig 3.7 with structural equations in Eq 3.11. And we present the corresponding causal DAG in Fig 3.6 for visualization of the assumed causal relation.

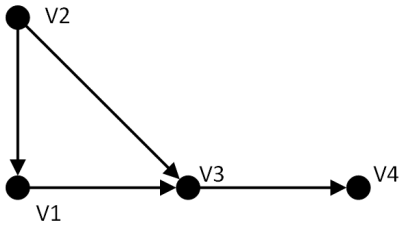


Figure 3.6: Assumed causal DAG

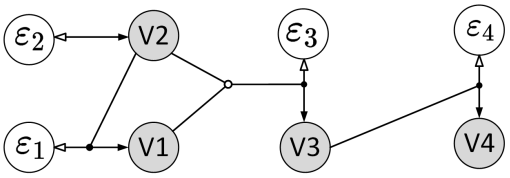


Figure 3.7: Linear causal structural model

$$\begin{cases} V_1 := f_1(\epsilon_1; V_2) = 3 \cdot V_2 + \epsilon_1 \\ V_2 := f_2(\epsilon_2) = \epsilon_2 \\ V_3 := f_3(\epsilon_3; V_1, V2) = 5 \cdot V_1 + 4 \cdot V_2 + \epsilon_3 \\ V_4 := f_4(\epsilon_4; V_3) = 6 \cdot V_3 + \epsilon_4 \end{cases} \tag{3.11}$$
$$\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \sim \mathcal{N}(0, 1)$$

In Fig 3.7, Each variable V_i is assigned with its unique exogenous noise ϵ_i. The bi-directional arrows indicate structural equations conditioned on its parent variable. Black and white arrowheads refer to the prediction and abduction directions respectively.

Associational Inference

We simulate an observational dataset according to the linear structural causal model in Fig 3.7. And we perform a linear regression of variables V_1 and V_3 according to dependence between them as in Fig 3.8. The coefficient in regression, 6.3, deviates from our assumption, 5.0. That is because the confounding effect of V_2 can distort the association between V_1 and V_3.

Interventional Inference

Next, we can use do-operation on the intervention level to protect the relationship between V_1 and V_3 from confounding of V_2. We intervene on V_1 with do-operation $do(V_1 := 0)$ and $do(V_1 := 1)$ for an example. Firstly, we replace structural equation $V_1 := f_1(\epsilon_1; V_2) = 3 \cdot V_2 + \epsilon_1$ with $V_1 := 0$ and $V_1 := 1$ respectively. This way, we block the causation from V_2 to V_1 and restrict V_1 to a constant. The modified causal DAG, Linear SCM, and structural equations are in Fig 3.9, 3.10, and Eq 3.12.

$$\begin{cases} V_1 := 1 \\ V_2 := f_2(\epsilon_2) = \epsilon_2 \\ V_3 := f_3(\epsilon_3; V_1, V2) = 5 \cdot V_1 + 4 \cdot V_2 + \epsilon_3 \\ V_4 := f_4(\epsilon_4; V_3) = 6 \cdot V_3 + \epsilon_4 \end{cases} \tag{3.12}$$
$$\epsilon_2, \epsilon_3, \epsilon_4 \sim \mathcal{N}(0, 1)$$

Then we compare the interventional inference with associational inference. We follow the modified SCM to generate post-interventional distribution $p(V_2, V_3|do(V_1 := 0))$ and $p(V_2, V_3|do(V_1 :=$

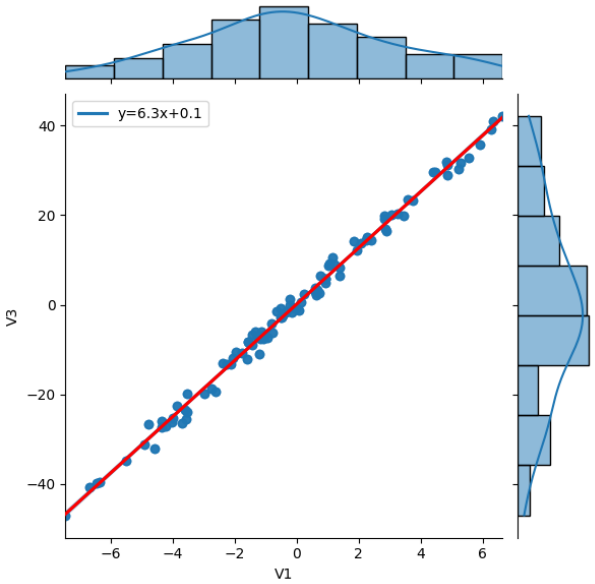


Figure 3.8: Linear regression of V_3 on V_1

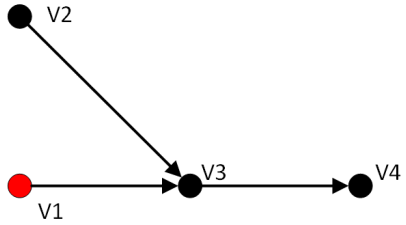


Figure 3.9: Causal DAG intervened on V_1

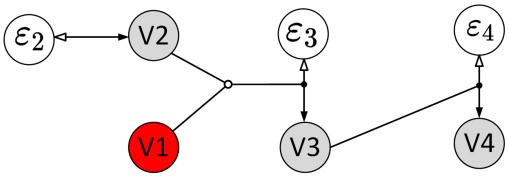


Figure 3.10: Linear causal structural model intervened on V_1

1)) in Fig 3.12. And we filter the observational dataset with $V_1 = 0$ and $V_1 = 1$ separately and plot conditional probability distribution $p(V_3, V_2|V_1 = 0)$ and $p(V_3, V_2|V_1 = 1)$ in Fig 3.12. In the observational distribution, the marginal distributions $p(V_2|V_1 = 0)$ and $p(V_2|V_1 = 1)$ are different. This is because the distribution of V_1 would affect the distribution of V_2 and finally results in a confounding effect on the distribution of V_3. In the interventional distribution, the marginal distributions $p(V_2|do(V_1 := 0))$ and $p(V_2|do(V_1 := 1))$ are similar. This is because V_1 has no causation on V_2, and the confounding path along $V_1 \to V_2 \to V_3$ is blocked. Thus the marginal distributions $p(V_3|do(V_1 := 0))$ and $p(V_3|do(V_1 := 1))$ could reflect the causal relation between V_3 and V_1 on the population level. This difference is also proved in section 3.2.2.

Counterfactual Inference

We take an example of counterfactual reasoning: for a specific observation v, what would its v_2 and v_3 be if its v_1 had been changed? Concretely, given a random individual v($v_1 = 0.0, v_2 = 0.02, v_3 = 1.91, v_4 = 11.20$) from observational data, infer its $\widetilde{v_2}$ and $\widetilde{v_3}$ after do($v_1 := 1$). We follow the three steps in counterfactual inference to answer this question.

- Abduction: We infer the exogenous noise ϵ from the individual observation by inverted structural equations as in Eq 3.13. The exogenous noise $p_{\mathfrak{G}}(\epsilon \mid \text{v})$ for sampled observation v is $\epsilon(\epsilon_1 = -0.05, \epsilon_2 = 0.02, \epsilon_3 = 1.84, \epsilon_4 = -0.28)$

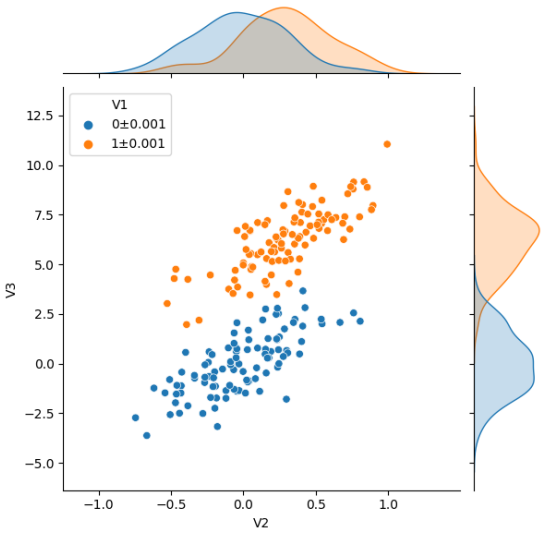


Figure 3.11: Distribution $p(V_2, V_3|V_1 = 0)$ and $p(V_2, V_4|V_1 = 1)$

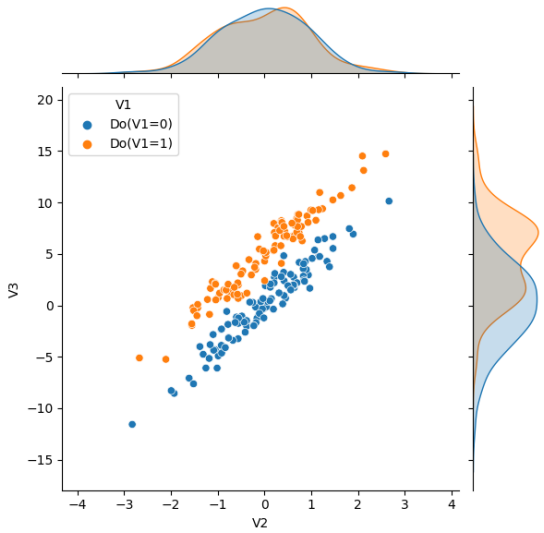


Figure 3.12: Distribution $p(V_2, V_3|do(V1 := 0))$ and $p(V_2, V_3|do(V_1 := 1))$

$$\begin{cases} \epsilon_1 := f_1^{-1}(v_1; v_2) = v_1 - 3 \cdot v_2 \\ \epsilon_2 := f_2^{-1}(v_2) = v_2 \\ \epsilon_3 := f_3^{-1}(v_3; v_1, v_2) = v_3 - 5 \cdot v_1 - 4 \cdot v_2 \\ \epsilon_4 := f_4^{-1}(v_4; v_3) = v_4 - 6 \cdot v_3 \end{cases} \tag{3.13}$$

- Action: We performed an intervention $\text{do}(v_1 := 1)$, resulting in a set of modified structural equations $\widetilde{\mathbf{S}}$. It combines with the exogenous noise $p_{\mathfrak{G}}(\epsilon \mid \text{v})$ and forms a modified SCM $\widetilde{\mathfrak{G}}$ as Eq 3.14. It has the same structural equations as Eq 3.12 but with individual exogenous noises rather than sampling noise from Gaussian distribution. The causal Bayesian Network and graph of SCM for the modified SCM $\widetilde{\mathfrak{G}}$ is the same as Fig 3.9 and Fig 3.10.

$$\begin{cases} V_1 := 1 \\ V_2 := f_2(\epsilon_2) = \epsilon_2 \\ V_3 := f_3(\epsilon_3; V_1, V2) = 5 \cdot V_1 + 4 \cdot V_2 + \epsilon_3 \\ V_4 := f_4(\epsilon_4; V_3) = 6 \cdot V_3 + \epsilon_4 \end{cases} \tag{3.14}$$
$$\epsilon(\epsilon_1 = -0.05, \epsilon_2 = 0.02, \epsilon_3 = 1.84, \epsilon_4 = -0.28)$$

- Prediction: We compute the value of interested variables based on the modified SCM, $p_{\widetilde{\mathfrak{G}}}(\text{v})$. We plot the factual individual and counterfactual individual on the interventional distribution $p(V_2, V_3|do(V_1 := 0))$ and $p(V_2, V_3|do(V_1 := 1))$ for visualization in Fig 3.13.

Finally, we can give the predicted quantity $\widetilde{\text{v}}(v_1 = 1.0, v_2 = 0.02, v_3 = 6.91, v_4 = 41.20)$ as our answer for the counterfactual reasoning. As the individual exogenous noises keep the same along the three steps, the implicit individual information compatible with the observation is preserved in counterfactual inference. Besides, comparing the factual and counterfactual inferred individuals in Fig 3.13, the v_2 remains the same. This match the fact that the intervention $do(v_1 := 1)$ protects v_1 from causation from v_2. In this way, we make a counterfactual inference on the individual level.

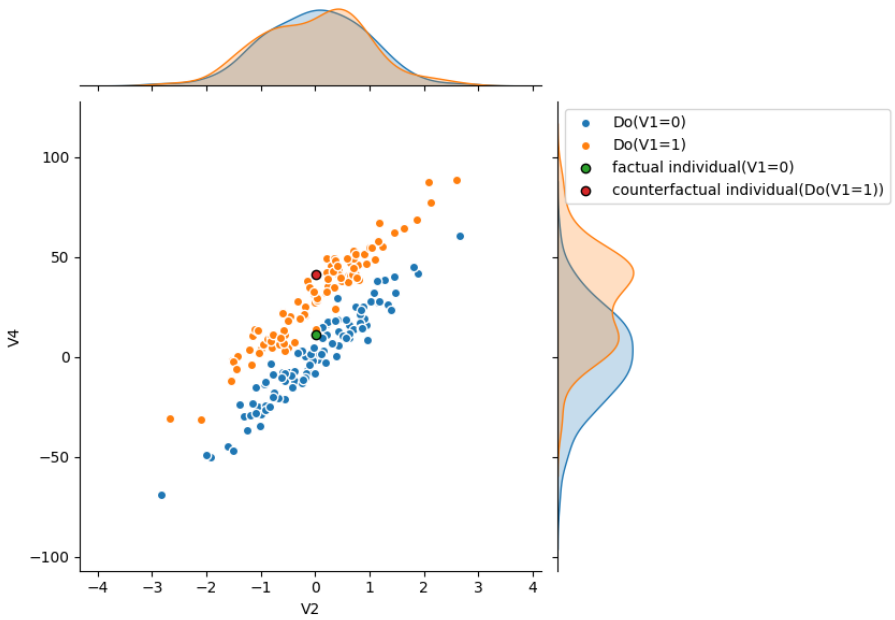


Figure 3.13: Counterfactual inference on a random individual

3.3 Generative Models for SCM

While SCM generally works well for low-dimensional scalar variables, it needs to be more flexible to model high-dimensional data such as images. We can replace linear structural equations with deep-learning techniques to overcome this limitation. And we call such structural equations deep structural equations.

Deep Learning is a field in AI that enables computational models composed of multiple processing layers to automatically learn representation from data[25]. When only the feature vectors of the data samples are available while the labels are missing, it is named unsupervised learning. Our task is classified into the probabilistic model estimation task of unsupervised learning. In the probabilistic model estimation task, for a probabilistic model that can generate samples, we use the samples to learn the statistical patterns in the data and parameters of the probabilistic model so that the samples generated by the probabilistic model are similar to the training samples. Furthermore, we focus on the image generative models.

In this section, we first introduce the normalizing flow and variational autoencoder in detail. After that, we introduce other deep generative models briefly.

3.3.1 Normalizing flow

Normalizing flow is a generative model that learns to transform a simple probability distribution into a more complex distribution. It is invertible and it models a bijective mapping between the two distributions. This characteristic matches the invertibility of structural equations between variables and exogenous noises. It was first proposed by Tabak and Vanden Eijnden[52]. Then it was applied by Dinh et al.[10] on density estimation, particularly on images. Later it was extended to conditional density estimation by Brian L. Trippe[53].

Within this section, we first elaborate on its transformation function, then discuss the application, density estimation and sampling. Finally, we introduce two examples of elementwise transformation: affine transformation and spline transformation.

Transformation function

Given a complex probability density $p\left(z^{(0)}\right), z^{(0)} \in \mathbb{R}^d$ and a simple known distribution $p\left(z^{(K)}\right)$, normalizing flow can apply a sequence of invertible transformation functions $f_1, \ldots, f_K : \mathbb{R}^d \to \mathbb{R}^d$ [21] to bijectively map from $p\left(z^{(0)}\right)$ to $p\left(z^{(K)}\right)$. These functions are required to be differentiable, invertible and, monotonic. And the likelihood of the input $z^{(0)}$ can be expressed as follows after change of variables:

$$p\left(z^{(0)}\right) = p\left(z^{(K)}\right) \cdot \prod_{k=1}^{K} \left| \det \frac{\partial f_k\left(z^{(k-1)}\right)}{\partial z^{(k-1)}} \right|, \tag{3.15}$$

where $z^{(k)} = f_k\left(z^{(k-1)}\right)$. The second term on the right is the determinant of the Jacobian for $f_1, \ldots, f_K$ and represents the volume change due to the transformations. This part ensures that the probability mass is normalized after each transformation.

Ideally, the normalizing flow can present any distribution $p\left(z^{(0)}\right)$ by a simple distribution such as Gaussian distribution if the transformations $f_1, \ldots, f_K$ are arbitrarily complex. However, the Jacobian computation of arbitrarily complex functions can be of high time complexity $O(n^3)$. So transformations are frequently designed to allow efficient computation of their Jacobian determinant. Recently popular transformations are those with Jacobian of a triangular matrix, such that the determinant of Jacobian could be simplified to be the multiplication of the Jacobian's diagonal, and the time complexity can decrease to $O(n)$:

$$\det \frac{\partial f_k\left(z^{(k-1)}\right)}{\partial z^{(k-1)}} = \prod_i \frac{\partial f_k\left(\boldsymbol{z}^{(k-1)}\right)_i}{\partial z_i^{(k-1)}} \tag{3.16}$$

Density estimation and sampling

Given observed data $\mathcal{D} = \{z_i\}_{i=1}^{M}$ from some unknown and complex distribution, we can optimize the parameters by maximizing the log-likelihood of the data samples:

$$\log p(\mathcal{D}; \theta) = \sum_{\boldsymbol{z}^{(0)} \in \mathcal{D}} \log p\left(\boldsymbol{z}^{(0)}; \boldsymbol{\theta}\right) = \sum_{\boldsymbol{z}^{(0)} \in \mathcal{D}} \left[\log p\left(\boldsymbol{z}^{(K)}\right) + \sum_{k=1}^{K} \log \left| \det \frac{\partial f_k\left(\boldsymbol{z}^{(k-1)}; \boldsymbol{\theta}_k\right)}{\partial \boldsymbol{z}^{(k-1)}} \right| \right] \tag{3.17}$$

Hence, a normalizing flow can be trained to model the dataset distribution via stochastic gradient descent. One advantage of normalizing flows is that it uses the exact likelihood as an objective without approximation.

With well trained parameters, a normalizing flow can estimate the probability density of object distribution $p\left(z^{(0)}\right)$ by transforming inversely the base distribution $p\left(z^{(K)}\right)$. And sampling is implemented by sampling from base distribution $p\left(z^{(K)}\right)$ and transforming inversely:

$$\begin{aligned} \tilde{z}^{(K)} &\sim p\left(z^{(K)}\right) \\ \tilde{z}^{(0)} &= f_1^{-1} \circ f_2^{-1} \circ \ldots \circ f_K^{-1}\left(\tilde{z}^{(K)}\right) \end{aligned} \tag{3.18}$$

Example: Affine transform

Affine transformation is the first used and simplest transformation class, also named location-scale transformation:

$$f_i\left(\boldsymbol{z}^{(i)}; \theta_i\right) = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{z}^{(i)} \text{ where } \quad \theta_i = \{\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i\}, \tag{3.19}$$

where $\boldsymbol{\sigma}_i$ controls the scale and $\boldsymbol{\mu}_i$ controls the location. The log absolute Jacobian determinant is:

$$\log \left| \det J_{f_i}(\boldsymbol{z}^{(i)}) \right| = \sum_{i=1}^{D} \log |\boldsymbol{\sigma}_i| \tag{3.20}$$

Although the affine transformation is simple to train, it has limited expressivity. For example, a single affine transformation of multivariate Gaussian results in a distribution whose conditional probability $p_{z'}\left(z'_{i+1} \mid \mathbf{z}'_i\right)$ will necessarily be Gaussian.

Example: Spline transform

Spline transformation is composed of piecewise simple low-order monotonic transformation functions rather than a sole monotonic transformation function. In each bin, the Jacobian determinant can be simple to compute and the overall transformation function can be more flexible than a sole monotonic transformation function. A popular type of spline transformation is rational-quadratic spline transform[11]. It divides the input domain and output domain into bijective bins and define a monotonically-increasing rational-quadratic transformation function at each bin. We give an instance of rational-quadratic spline transform in Fig 3.14.

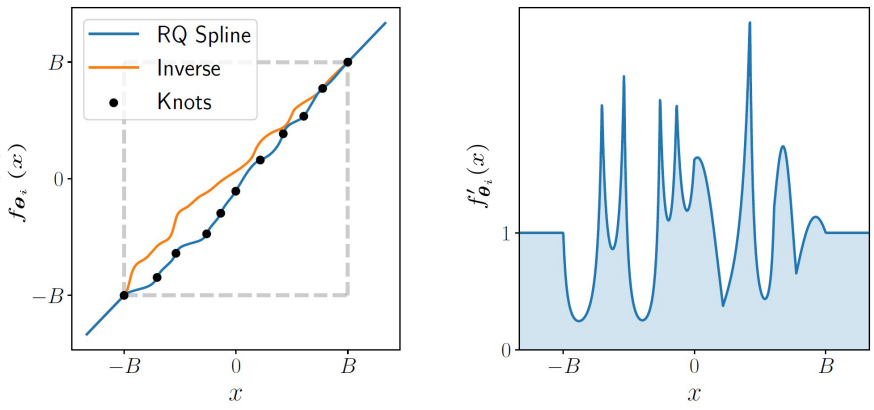


Figure 3.14: Example of a rational-quadratic spline transform. Image from [11]

Firstly, we define the parameters for the monotonic rational-quadratic spline transformation, the derivatives and the location of each knot. Given an input domain $[-B, B]$ and an output domain $[-B, B]$, we define $K+1$ knots $\left\{\left(x^{(k)}, y^{(k)}\right)\right\}_{k=0}^{K}$ to segment the transformation function to K different rational-quadratic transformation functions. These knots are monotonically increasing and we set boundary as $\left(x^{(0)}, y^{(0)}\right) = (-B, -B)$ and $\left(x^{(K)}, y^{(K)}\right) = (B, B)$. Then we define the derivatives at these knots as $\left\{d^{(k)} > 0\right\}_{k=0}^{K+1}$. To generalize the transformation to unconstrained input and output, we set the transformation outside the domain $[-B, B]$ as identity by setting the boundary derivatives $\delta^{(0)} = \delta^{(K)} = 1$.

In the k^{th} bin, we define the rational quadratic transformation function $f_{\boldsymbol{\theta}_i}^{(k)}(\xi)$ that passes the neighboring knots and have the corresponding gradients at knots as Eq 3.3.1.

$$f_{\boldsymbol{\theta}_i}^{(k)}(\xi) = y^{(k)} + \frac{\left(y^{(k+1)} - y^{(k)}\right)\left[s^{(k)}\xi^2 + \delta^{(k)}\xi(1-\xi)\right]}{s^{(k)} + \left[\delta^{(k+1)} + \delta^{(k)} - 2s^{(k)}\right]\xi(1-\xi)} \tag{3.21}$$

where $s_k = \left(y^{k+1} - y^k\right) / \left(x^{k+1} - x^k\right)$ and $\xi(x) = \left(x - x^k\right) / \left(x^{k+1} - x^k\right)$.

Note the rational quadratic transformation function $f_{\boldsymbol{\theta}_i}^{(k)}(\xi)$ on the k^{th} bin is monotonically increasing as its derivative with respect to input x is non-negative. The derivative is shown in Eq 3.3.1.

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[f_{\boldsymbol{\theta}_i}^{(k)}(\xi)\right] = \frac{\left(s^{(k)}\right)^2\left[\delta^{(k+1)}\xi^2 + 2s^{(k)}\xi(1-\xi) + \delta^{(k)}(1-\xi)^2\right]}{\left[s^{(k)} + \left[\delta^{(k+1)} + \delta^{(k)} - 2s^{(k)}\right]\xi(1-\xi)\right]^2} \tag{3.22}$$

So far, we define a rational quadratic transformation function at each bin. At each bin, the knots are monotonical increasing and the gradient of each knot is non-negative. We combines these rational quadratic transformation functions $f_{\boldsymbol{\theta}_i}^{(k)}(\xi)$ to an invertible and monotonic rational-quadratic spline transformation function $f_{\boldsymbol{\theta}_i}$. And we compute the logarithm of the absolute value of the determinant of its Jacobian as the sum of the logarithm of the derivatives in Eq. 3.3.1. It shows in Eq 3.23.

$$\begin{aligned}
\log\left|\det f_{\boldsymbol{\theta}_i}(x)\right| &= \sum_{k=0}^{K}\sum_{x\in[x^k,x^{k+1}]}\log\left|\det f_{\boldsymbol{\theta}_i}^{(k)}(x)\right| \\
&= \sum_{k=0}^{K}\sum_{x\in[x^k,x^{k+1}]}\log\frac{\mathrm{d}}{\mathrm{d}x}\left[f_{\boldsymbol{\theta}_i}^{(k)}(\xi)\right] \\
&= \sum_{k=0}^{K}\sum_{x\in[x^k,x^{k+1}]}\log\frac{\left(s^{(k)}\right)^2\left[\delta^{(k+1)}\xi^2 + 2s^{(k)}\xi(1-\xi) + \delta^{(k)}(1-\xi)^2\right]}{\left[s^{(k)} + \left[\delta^{(k+1)} + \delta^{(k)} - 2s^{(k)}\right]\xi(1-\xi)\right]^2}
\end{aligned} \tag{3.23}$$

And we can inverse rational quadratic transformation function $f_{\boldsymbol{\theta}_i}^{(k)}(\xi)$ as Eq 3.24.

$$\begin{aligned}
x &= \xi^{-1}(x^{k+1}, x^k) = \xi(x)\left(x^{k+1} - x^k\right) + x^k \\
\xi(x) &= 2c/\left(-b - \sqrt{b^2 - 4ac}\right) \\
a &= \left(y^{(k+1)} - y^{(k)}\right)\left[s^{(k)} - \delta^{(k)}\right] + \left(y - y^{(k)}\right)\left[\delta^{(k+1)} + \delta^{(k)} - 2s^{(k)}\right] \\
b &= \left(y^{(k+1)} - y^{(k)}\right)\delta^{(k)} - \left(y - y^{(k)}\right)\left[\delta^{(k+1)} + \delta^{(k)} - 2s^{(k)}\right] \\
c &= -s^{(k)}\left(y - y^{(k)}\right)
\end{aligned} \tag{3.24}$$

3.3.2 Variational Autoencoder

Autoencoder is traditionally used for dimensionality reduction and representation learning[44]. More recently, however, theoretical advancement in latent variable models has resulted in the variational autoencoder (VAE) [20], which is primarily used as a generative model.

Autoencoder maps input to a representation in low dimension and then reconstructs the input. It is trained by minimizing the difference between the original and reconstructed data. Instead of being encoded into a deterministic representation in Autoencoder, the data is mapped to a probability distribution function in VAE. The input is then reconstructed by sampling from this distribution. By this means, the VAE achieve density estimation and sampling high-dimensional data with limited latent space.

Within the following sections, we first briefly introduce Autoencoder as a base. Then we discuss VAE on the problem formulation, model structure, loss function, and the parametrization trick in training.

Autoencoder

Autoencoder was first introduced by H. Bourlard in 1988[2], and is used for dimensionality reduction or representation learning since then.

The object of an autoencoder is to reconstruct the input data as similar as possible. As illustrated in Fig 3.15, it is made up of an encoder and a decoder, both of which are neural

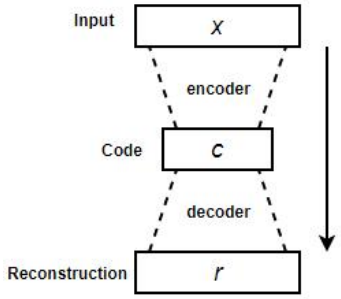


Figure 3.15: Autoencoder Structure

networks. Given a data sample x as an input, the encoder will map it into a deterministic representation c in a low dimension. The decoder then attempt to transform the representation c to a reconstruction r. The task is to make the reconstruction r similar to the input x based on some difference measure. In this way, the autoencoder expects to learn informative properties of the input x in a low representation dimension.

Problem Formulation for Variational Autoencoder

Although sharing a similar encoder and decoder structure, VAE is designed for a different problem. To understand the VAE well, we first give a clear problem formulation.

Given a data set $X = \{x^{(1)}, \ldots, x^{(N)}\}$, we assume it consists of N Independent and identically distributed samples from a variable x and it is generated from a random latent variable z. We assume the latent variable z is samples from a prior distribution $p_{\boldsymbol{\theta}^*}(\boldsymbol{z})$. And the data set $X = \{x^{(1)}, \ldots, x^{(N)}\}$ is generated from a conditional distribution $p_{\boldsymbol{\theta}^*}(\boldsymbol{x} \mid \boldsymbol{z})$. Importantly, we cannot know but only approximate the true parameters θ^* and the true latent variables value z.

To estimate the probability density of the data and describe the relationships between the data and the latent variables. We have to compute the marginal likelihood $p_{\boldsymbol{\theta}}(x) = \int p_{\boldsymbol{\theta}}(\boldsymbol{z}) p_{\boldsymbol{\theta}}(x \mid z) dz$, and the true posterior density $p_{\boldsymbol{\theta}}(\boldsymbol{z} \mid \boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) / p_{\boldsymbol{\theta}}(\boldsymbol{x})$. However, the integral is intractable as the $p_{\boldsymbol{\theta}}(x \mid z)$ is too complex to integrate. The posterior is also intractable as a result of intractable $p_{\boldsymbol{\theta}}(x)$. VAE adopts an approximation $q_\phi(z \mid x)$ for the intractable true posterior $p_{\boldsymbol{\theta}}(z \mid x)$ [20]. Then the task for VAE becomes to learn the posterior approximation model parameters ϕ and the generative model parameters θ with neural networks.

Variational Autoencoder Architecture

We can use an encoder to learn the $q_\phi(z \mid x)$ and a decoder to learn $p_{\boldsymbol{\theta}}(\boldsymbol{x} \mid \boldsymbol{z})$ in VAE. However, unlike autoencoder, which directly takes the code c as latent variables, VAE takes a latent distribution and sample latent variables z from this distribution. So in the flow of VAE, the encoder predicts a distribution over the possible representation z that we assume x has been generated from, Then VAE samples from this distribution. Next, the decoder takes the samples as input to produce a distribution over the possible x. The structure of VAE is illustrated in Fig 3.16.

We can take any neural network as an encoder and decoder. Specifically for image data, it is common to use the convolutional layer[26] and the corresponding deconvolutional layer as the encoder and decoder, respectively.

Loss Function: the Evidence Lower Bound

D. P. Kingma & M. Welling[20] describe a suitable objective function to train the variational autoencoder. It is the logarithm of the marginal likelihood of the input data $\log p_{\boldsymbol{\theta}}(\boldsymbol{X})$. And it is

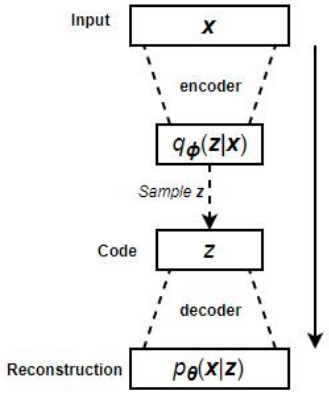


Figure 3.16: Variational Autoencoder Structure

equal to the sum over the marginal likelihoods of each data sample as they are all assumed to be independent and identically distributed.

$$\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\right) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) \tag{3.25}$$

Each of the term of the sum on the right side can be rewritten as:

$$\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) = D_{KL}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right) \| p_{\boldsymbol{\theta}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)\right) + \mathcal{L}\left(\boldsymbol{\theta}, \phi; \boldsymbol{x}^{(i)}\right) \tag{3.26}$$

where D_{KL} denotes Kullback-Leibler (KL) Divergence [22] presenting the difference between two probability distributions. In this equation it approximate the difference between approximate posterior and the true posterior. $\mathcal{L}\left(\boldsymbol{\theta}, \phi; \boldsymbol{x}^{(i)}\right)$ denotes the evidence lower bound(ELBO) on the marginal likelihood $p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right)$.

As the KL Divergence is non-negative, We take the ELBO as a lower bound for $\log p_{\boldsymbol{\theta}}(\boldsymbol{X})$ and maximize the ELBO to indirectly maximize $\log p_{\boldsymbol{\theta}}(\boldsymbol{X})$. The ELBO can be formulated as Eq 3.27.

$$\begin{aligned} \mathcal{L}\left(\boldsymbol{\theta}, \phi; \boldsymbol{x}^{(i)}\right) &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[-\log q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) + \log p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z})\right] \\ &= -D_{KL}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right) \| p_{\boldsymbol{\theta}}(\boldsymbol{z})\right) + \mathbb{E}_{q_\phi\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right] \end{aligned} \tag{3.27}$$

The $D_{KL}\left(q_\phi\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right) \| p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)$ is the KL Divergence of the approximate posterior $q_\phi\left(\boldsymbol{z} \mid x^{(i)}\right)$ and the prior $p_{\boldsymbol{\theta}}(\boldsymbol{z})$. It is a regulation guiding the learned approximation $q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid x^{(i)}\right)$ (the encoder) to match the chosen prior $p_{\boldsymbol{\theta}}(\boldsymbol{z})$. In this way, the model can be robust to small perturbations along the latent distribution. Namely, the encoder can be general and less likely to overfit. Conventionally, prior distribution is assumed as the standard Gaussian distribution, and approximate posterior is assumed as the independent multivariate Gaussian distribution.

The $\mathbb{E}_{q_\phi\left(\boldsymbol{z}|\boldsymbol{x}^{(i)}\right)}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right]$ is the probability density of generated output given the predicted latent distribution over z. Its negative can be used to measure reconstruction error. It describes how accurately the output reconstructs the input. For image data, it usually takes mean squared error (MSE).

By differentiating and maximizing the ELBO with respect to both the variational parameters ϕ and the generative parameters θ, we can train a model with a trade-off between expressiveness and conciseness. The model is expected to reconstruct the input data while learning a simple latent distribution close to the prior distribution. In practice, we usually take the negative ELBO as a loss function and minimize it in practice.

Training the Variational Autoencoder

Take training a VAE for image data as an example. We set the prior $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ as a standard Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$, and set the approximate posterior $q_\phi\left(z \mid x^{(i)}\right)$ as an independent multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}))$ with $\boldsymbol{\mu} \in \mathbf{R}^k$ and $\boldsymbol{\sigma} \in \mathbf{R}^k$. k is the defined latent dimension. Together with the above loss function, we can define a complete variational autoencoder architecture as Fig 3.17.

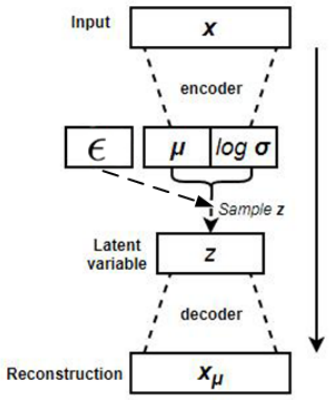


Figure 3.17: Variational Autoencoder Structure with multivariate Guassian prior and posterior

The VAE takes an image data x as input, and transforms it through a neural network(encoder) into parameters μ and $\log \sigma$ that describe the approximate posterior $q_\phi\left(z \mid x^{(i)}\right)$ (where $\phi = (\mu, \sigma)$). Then, we have to sample a latent value z from the distribution $\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}))$. However, such a sampling process in the model will make the model non-deterministic and non-differentiable with respect to its learned parameters. We implement a parametrization trick to solve the problem by moving the sampling process outside the model. We set additionally auxiliary random variables $\epsilon \sim \mathcal{N}(0, 1)$ as input and replace sampling $\boldsymbol{z}$ from $\mathcal{N}(\boldsymbol{\mu}, \text{diag}(\sigma))$ by computing $z = \mu + \sigma \odot \epsilon$ (where $\odot$ denotes the element-wise product). Finally, this z is propagated through a decoder network (conventionally the inverse architecture of the encoder) into a reconstruction $\hat{x}$ of the input x. we use the differece between $\hat{x}$ and x to evaluate the reconstruction error via MSE.

Finally, we construct a generative model with input x and ϵ and its loss function that can be differentiated with respect to its learned parameters. Then, we can train this model using backpropagation with stochastic gradient descent.

3.3.3 Other Generative Models

We briefly introduce Generative adversarial network and Diffusion model in this section.

Generative adversarial network

Generative adversarial network(GAN) is an unsupervised deep learning model which uses two neural networks to compete with each other[15]. The GAN model architecture involves two submodels: a generator model to generate new examples and a discriminator model to classify whether the generated examples are the observational data or the generated data. The structure of GAN is shown in Fig 3.18.

We train a GAN as follows. Given a generator model $G : \mathbb{R}^D \mapsto \mathbb{R}^K$, with parameters Θ, a discriminator $D : \mathbb{R}^K \mapsto [0, 1]$, with parameters Φ. We take cross entropy as the loss function for GAN as Eq 3.28. In practice, we use Monte Carlo to estimate the cross entropy and alternatively update the model parameters Θ and Φ by $\min_\Theta$ and $\max_\Phi$.

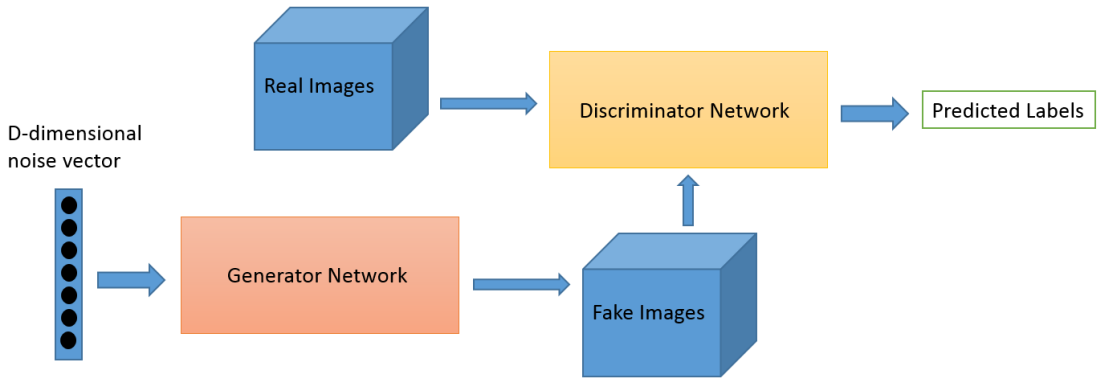


Figure 3.18: The structure of Generative adversarial network

$$\min_{\Theta} \max_{\Phi} \mathbb{E}_{\boldsymbol{x}\sim p^*}[\log D(\boldsymbol{x})] + \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}[\log(1 - D(G(\boldsymbol{\epsilon})))], \tag{3.28}$$

where p^* is the true distribution, ϵ is a Gaussian noise.

Diffusion

Diffusion model is a generative model which lets observational data diffuse through multiple latent space layers[18]. It defines a forward diffusion process mapping observational data distribution into Gaussian distribution by adding a small amount of Gaussian noise to the observational data distribution step by step. Then it reverses the forward diffusion process to learn the map from the Gaussian distribution into the observational data distribution. The map in each step can be treated as a VAE, and we can optimize the log-likelihood by maximizing ELBO. The structure of Diffusion is in Fig 3.19.

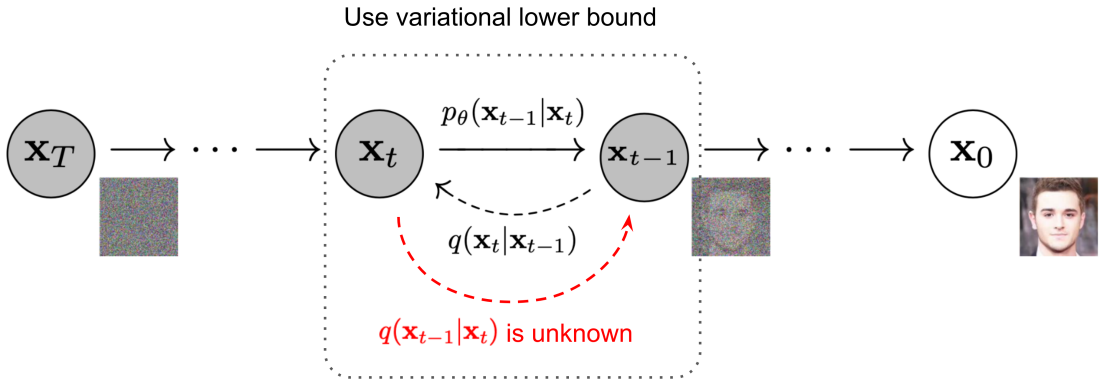


Figure 3.19: The structure of Diffusion model. Image from [18]

Firstly, it defines a forward diffusion process as follows. Given a data sample from an observational data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, we add a little Gaussian noise into the sample for each step in T steps like Eq 3.29 and produce a series of latent space $\mathbf{x}_1, \ldots, \mathbf{x}_T$. Eventually, when $T \to \infty$, $\mathbf{x}_T$ is equivalent to a Gaussian distribution.

$$q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)=\mathcal{N}\left(\mathbf{x}_t ; \sqrt{1-\beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \quad q\left(\mathbf{x}_{1: T} \mid \mathbf{x}_0\right)=\prod_{t=1}^T q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) \tag{3.29}$$

Next, we inverse the forward diffusion process and use $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ to estimate unknown $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$. We can treat a separate diffusion step between X_t and X_{t-1} as a VAE, and we can optimize the log-likelihood of observational data $\log p_\theta\left(\mathbf{x}_0\right)$ by maximizing ELBO.

Finally, we compose the diffusion steps into a compact model as Eq 3.30.

$$p_\theta\left(\mathbf{x}_{0: T}\right)=p\left(\mathbf{x}_T\right) \prod_{t=1}^T p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \tag{3.30}$$

Chapter 4

Related work

In this section, we summarize the causal models supporting counterfactual inference on images. They all follow the structural causal model(SCM) (see definition 3.2.5) and comply with the three steps in counterfactual inference(see definition 3.2.6). However, they use different deep learning architectures to model the causation from the attributes to the images. We include three state-of-art causal models supporting counterfactual inference on images, deep structural causal model (DSCM), ImageCFGen, and Diff-SCM. In the following content, we mainly elaborate on their architecture, counterfactual inference implementation and performance.

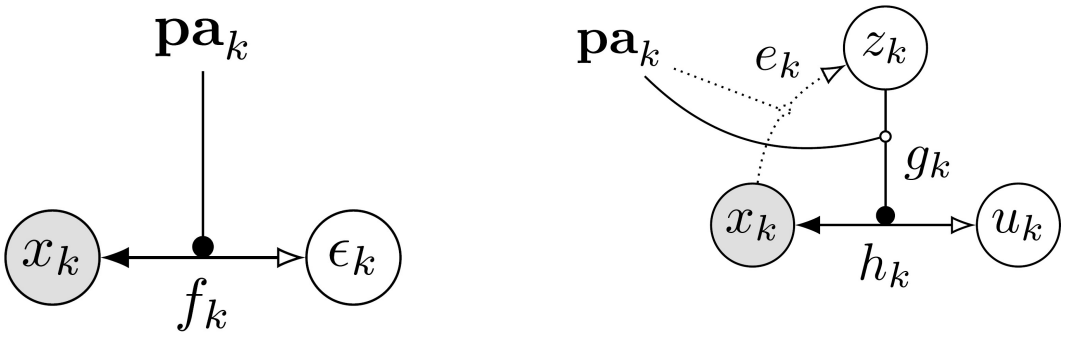
4.1 Deep Structural Causal Model

Deep structural causal model (DSCM) [37] adopts normalizing flow and variational inference to learn a deep structural equation f_k for each variable, $\mathbf{x_k} := f_k\left(\epsilon_k; \mathbf{pa}\left(\mathbf{x_k}\right)\right)$. Concretely, for a low-dimensional $\mathbf{x_k}$, DSCM learns an conditional normalizing flow [56] whose parameters are predicted by a neural network with the input of its parent variable $\mathbf{pa}_k$ as in Fig 4.3(a). As for a high dimensional variable $\mathbf{x_k}$ like images, DSCM adopts a amortized and explicit mechanism as in Fig 4.3(b). It includes a shallow invertible transformation h_k mapping from the variable $\mathbf{x_k}$ to its exogenous noise u_k. Besides, it has a deep non-invertible transformation from the variable $\mathbf{x_k}$ to the parameters in h_k. In the deep non-invertible transformation, DSCM adopts variational inference [20] e_k to approximate the other exogenous noise z_k, and the z_k is used by a decoder g_k to predict the parameters in h_k. The deep non-invertible transformation is designed to capture the high-level structure of the image data and reflected it in the predicted parameters in h_k. These deep structural causal equations form a deep structural causal model $\mathfrak{G} = (\mathbf{S}, \mathbf{p}(\epsilon))$,

Given an instance $x(x_0, ..., x_K)$, where $x_k, k \in K$ is the observed variables, the DSCM implements counterfactual inference as follows.

Abduction

For a low dimensional observed variable x_k, DSCM inverts the learned normalizing flow to map the observed variable x_k to its exogenous noise, $\epsilon_k = f_k^{-1}\left(x_k; \mathbf{pa}(x_k)\right)$. Moreover, for a high dimensional observed variable image x_k, DSCM first infers the first exogenous noise z_k by the encoder, $z_k = e_k(x_k, \mathbf{pa}_k)$. Then DSCM predicts the parameters in h_k by the decoder, $g_k(z_k, \mathbf{pa}_k)$. With the predicted parameters, the shallow invertible transformation h_k is compact and we infer the second exogenous noise u_k by the inverted normalizing flow, $u_k = h_k^{-1}(x_k; g_k(z_k, \mathbf{pa}_k))$. The abduction step follows the white arrowheads in Fig 4.3. As a result, DSCM obtains exogenous noises, $p_{\mathfrak{G}}(\epsilon \mid x)$



(a) structural equation for a low dimensional variable

(b) structural equation for high dimensional variable, a.k.a amortized and explicit mechanism

Figure 4.3: Two types of deep structural equation. Bi-directional arrows denote invertible transformations. It conditions on other inputs with edges ending in black circles. Black and white arrowheads refer to the prediction and abduction directions respectively. Dotted arrows denote an amortized variational approximation. Image from [37]

Action

DSCM intervenes on the causal DAG by directly reassigning the desired value to a variable. In this way, the intervened variable is deterministic. It isolates itself from its parent variables and can only affect its child variables. As a result, DSCM obtains a set of intervened deep structural equations $\widetilde{\mathbf{S}}$.

Prediction

A modified DSCM is formed after abduction and action, $\widetilde{\mathfrak{G}} = \left(\widetilde{\mathbf{S}}, p_{\mathfrak{G}}(\epsilon \mid x)\right)$. Then the counterfactual variables $\widetilde{x}$ can be computed by the modified DSCM $\widetilde{\mathfrak{G}}$. The prediction step follows the black arrowheads in Fig 4.3.

DSCM has been validated on a synthetic MNIST dataset, and a brain MRI scans medical dataset in gray scale. It has a good performance of counterfactual inference on luminosity and shape. Note that as there is no valid normalizing flow for the discrete variables, DSCM uses an alternative method Gumbel-Max distribution to estimate the distribution of the discrete variable.

4.2 ImageCFGen

ImageCFGen [8] adopts the GAN architecture with an additional variational inference for the abduction step. Given an image $\mathbf{x}$ and a set of its attributes a, ImageCFGen separates the SCM of the attributes a out of the complete SCM and applies the same invertible transformations in Fig 4.3 for it. For the high-dimensional image $\mathbf{x}$, it combines with the attributes a and passes through an encoder E to infer exogenous noises z. Then it combines with the attributes after intervention a_c and passes to a generator G to obtain a counterfactual inferred image $\mathbf{x_c}$. Meanwhile, it reconstructs the input image $\mathbf{x_r}$ in the similar workflow but not intervening on the attributes a. Then the counterfactual inferred image $\mathbf{x_c}$ and reconstructed image $\mathbf{x_r}$ are discriminated by a classifier. We present the structure of ImageCFGen in Fig 4.4. The ImageCFGen is trained with a loss function of cross entropy as Eq 3.28. And its counterfactual inference follows the upper part workflow in Fig 4.4.

ImageCFGen has a state-of-art performance of counterfactual inference on the MNIST dataset and a more complex CelebA dataset. It outperforms DSCM in generating high-quality valid

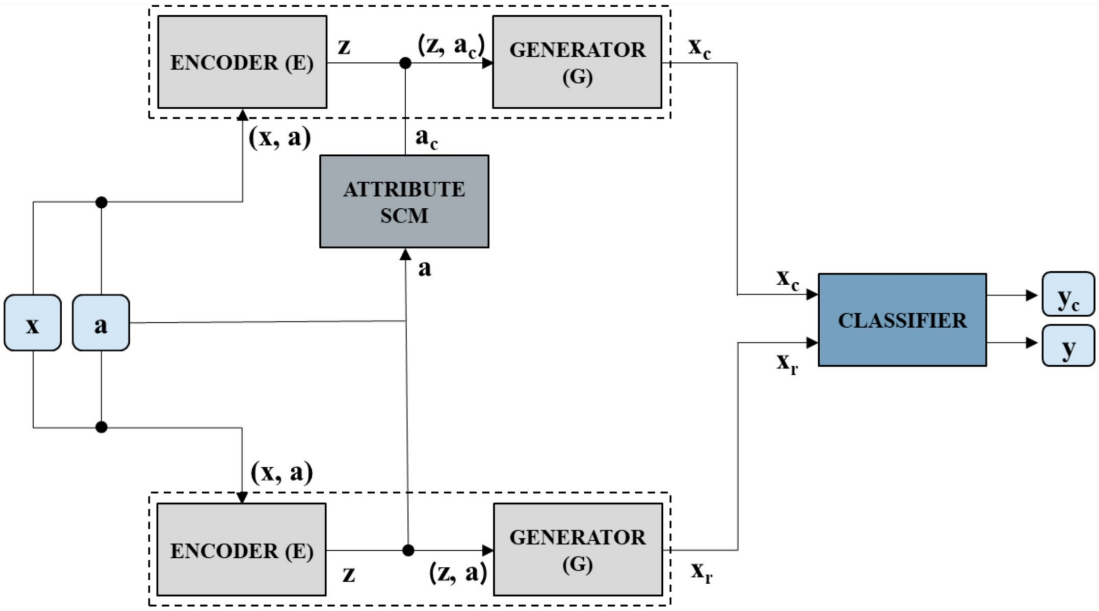


Figure 4.4: The structure of ImageCFGen. Image from [8]

counterfactual inferred images in the CelebA dataset.

4.3 Diff-SCM

Diff-SCM [9] adopts the Diffusion model architecture with a classifier guidance framework to guide the prediction toward the counterfactual distribution. The forward diffusion process of Diff-SCM is shown in Fig 4.5.

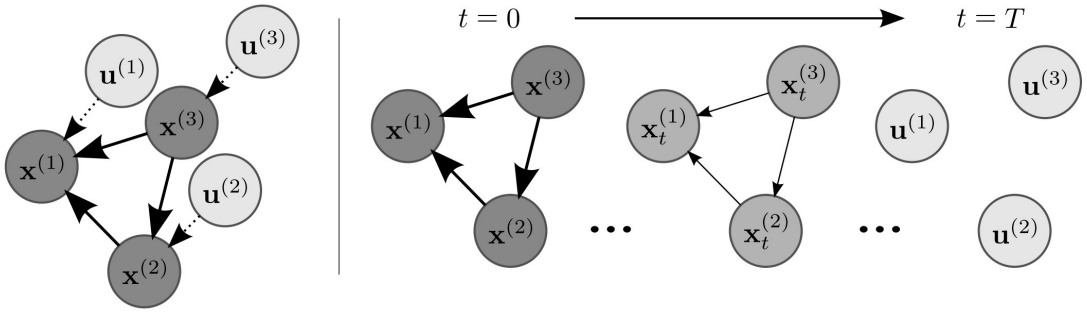


Figure 4.5: A forward diffusion process. Left: an example SCM with variables $\mathbf{x}^{(k)}$ and corresponding exogenous noises $\mathbf{u}^{(k)}$. Right: A forward diffusion process from variables $\mathbf{x}$ to their exogenous noises $\mathbf{u}$. Image from [46]

.

Given an SCM $\mathfrak{G}$, Diff-SCM uses the forward diffusion process to encode each variable $\mathbf{x}^{(k)}$ to its exogenous noise $\mathbf{u}^{(k)}$ as the abduction step. It can be considered as a process to weaken the causal relationships between variables $\mathbf{x}^{(k)}$ until they become independent at $t = T$. Meanwhile, the reverse diffusion process can be considered as a reconstruction step that reconstructs the causal relationships between variables $\mathbf{x}^{(k)}$. Based on these process, Diff-SCM takes an anti-causal predictor to guide the reverse diffusion process deviating from the reconstruction and toward the counterfactual distribution. This guided reverse diffusion process achieves the prediction under intervention.

The anti-causal predictor adopts the idea of the classifier guidance [9] and is defined as follows. Given an intervention on variable $\mathbf{x}^{(j)}$, $\text{do}(\mathbf{x}^{(j)} := x^{(j)})$, and its child variable $\mathbf{x}^{(i)}$, the anti-causal predictor is a neural network classifier to predict $p_{\widetilde{\mathfrak{G}}}\left(\mathbf{x}^{(j)} := x^{(j)} \mid x_t^{(i)}\right), t \in T$. Then the gradient of the anti-causal predictor with respect to the child variable $\nabla_{x_t^{(i)}} p_{\widetilde{\mathfrak{G}}}\left(\mathbf{x}^{(j)} := x^{(j)} \mid x_t^{(i)}\right)$, is weighed by a hyperparameter s and is subtracted from the sampled Gaussian noise in the reverse diffusion process. In this way, the anti-causal predictor guides the reverse diffusion process to the counterfactual distribution.

Diff-SCM has a state-of-art performance of counterfactual inference with a bi-variable causal model on the MNIST dataset and ImageNet dataset. However, it cannot be applied to a more complex causal model as the anti-causal predictor will be biased due to confounding. For example in Fig 4.5, the anti-causal predictor $p_{\widetilde{\mathfrak{G}}}\left(\mathbf{x}^{(2)} := x^{(2)} \mid x_t^{(1)}\right), t \in T$ involves the confounding from $\mathbf{x}^{(3)}$ and so is its gradient with respect to $\mathbf{x}^{(1)}$. As a result, the inferred counterfactual distribution is biased due to the confounding. Because Diff-SCM can not handle the confounding issue well and confounding is the core limitation to mitigate in our task, we don't adopt Diff-SCM in our project.

Chapter 5

Materials and Methods

This chapter elaborates on the Maastricht study dataset and methods adopted for this project.

5.1 Materials

Maastricht Study is an observational prospective population-based cohort study on type-2 diabetes[49]. From Maastricht Study, we extract a dataset consisting of participants with normal glucose metabolism and with type-2 diabetes. It records patient attributes, age, sex, type-2 status, and several fundus images at one visit for each participant. We split the dataset as Table 5.1. Participants' age ranges from 40 to 80 years old.

Table 5.1: Number of images and participants for the train/validation/test for Maastricht Study

Dataset	Images	Pariticipants	Gender		Type-2 diabetes status	
Train	13001	1796	All	1796	All	1796
			Male	849	Normal	1328
			Female	947	type-2 diabetes	468
Validation	2976	409	All	409	All	409
			Male	195	Normal	297
			Female	214	type-2 diabetes	112
Test	5923	819	All	819	All	819
			Male	373	Normal	626
			Female	446	type-2 diabetes	193

5.2 Methods

This section will explain how we construct a custom deep structural causal model(DSCM) on fundus images. We first discuss our assumptions. Next, we define deep structural equations for patient attributes and fundus images. After that, we elaborate on how to implement the three counterfactual inference steps. Finally, we introduce the image preprocessing methods adopted in our project.

5.2.1 Assumptions

We assume a causal Bayesian Network for age a, gender g, type-2 diabetes d, and fundus images i as Fig 5.1. And we assume there is no unobserved confounding.

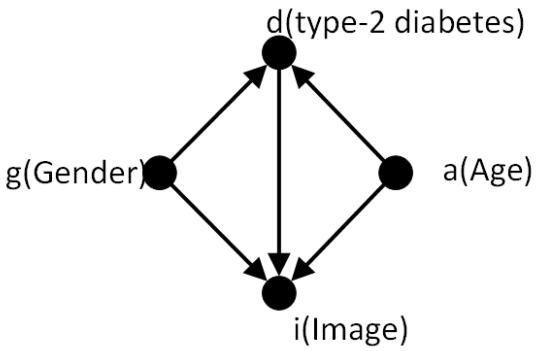


Figure 5.1: causal Bayesian Network for age, gender, type-2 diabetes, and fundus images

As shown above, we consider gender and age as confounders on type-2 diabetes and fundus images. This is based on the following research findings about their associations. The deep learning has successfully predicted type-2 diabetes with fundus images[17]. This implies an association between type-2 diabetes and fundus images. Besides, deep learning based on fundus images also predicts age and gender with high accuracy[38]. This implies associations between age, gender and fundus images. Moreover, it is stated in section 3.1 that older people and males are more likely to have type-2 diabetes[14, 4]. So there should be associations between age, gender and type-2 diabetes. As gender and age are not caused by any factor in the common sense, we assume the causal association should be from age and gender towards type-2 diabetes and fundus images. As a result, the gender and age are confounders in our assumption.

5.2.2 Deep structural equations

As the exogenous noises are mutually independent in definition 3.2.5 and the structural equations are under the Causal Edges Assumption in assumption 3.2.3, the structural equations are independent of each other. So we define a deep structural equation for each patient attribute and train them separately. In particular, we define an amortized explicit deep structural equation with variational autoencoder and normalizing flow adopted for fundus images because of its high dimension. This design will be explained in section 5.2.2 in detail. These deep structural equations finally form a custom overall deep structural causal model as in Fig 5.2.

Gender

For patient attribute gender g, it is a binary variable and has no parent variables in the causal DAG. We assume it follows the Bernoulli distribution, and we optimize its parameter probability by maximizing the likelihood of the observational data in the Bernoulli distribution.

Age

For patient attribute age a, it is a discrete variable with no parent variables in the causal DAG. We design a normalizing flow as its deep structural equation in Eq 5.1. Considering the variable age a is distributed over a constrained range, the normalizing flow first learns the spline flow in unconstrained space and then maps to the original range $[40, 80]$ using fixed affine transformation and exponential transformation.

$$a := f_A(\epsilon_A) = (\exp \circ \text{ AffineNormalization } \circ \text{ Spline }_\theta)(\epsilon_A), \tag{5.1}$$

where ϵ_A denotes an exogenous noise for a and follows Gaussian distribution $\epsilon_A \sim N(0, 1)$, Spline $_\theta$ denotes a rational-quadratic neural spline flow, AffineNormalization denotes an affine transforma-

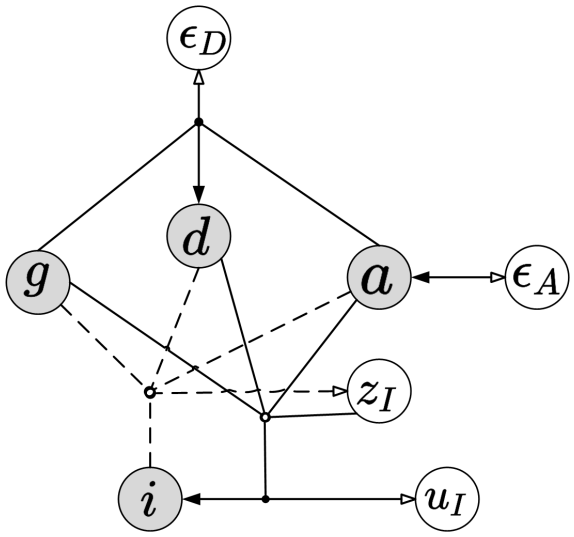


Figure 5.2: Custom deep structural causal model on fundus images. Bi-directional arrows indicate deep structural equations. It is conditioned on other inputs when it includes edges ending on itself in black circles. Black arrowheads denotes generative direction and white arrowheads denotes abductive directions. And dotted arrows depict an amortized variational approximation.

tion $f(\boldsymbol{x}) = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{x}$, where $\boldsymbol{\mu} = \frac{\sum_{n=0}^{N} \log a^{(n)}}{N}, \sigma = \sqrt{\frac{\sum \left(\log a^{(n)} - \mu\right)^2}{N}}$, exp denotes an exponential transformation $f(x) = e^x$.

We maximize the log-likelihood of the observational data to optimize the parameters, location and gradients of knots, for the rational-quadratic neural spline flow as Eq 3.17.

Type-2 diabetes

For patient attribute type-2 diabetes d, it is a binary variable and a child variable of gender and age. Considering Gumbel–max parametrization on such discrete variable has been proved with necessary properties for counterfactual[34] like invariance to category order and counterfactual stability. Counterfactual stability means that the model cannot produce a different counterfactual outcome unless the odds of the observed outcome increase less than other outcomes after the intervention. Namely, only when the probability of an alternative outcome has increased faster than the probability of the observed outcome after the intervention can the model possibly produce the alternative outcome as the counterfactual outcome. The Gumbel-Max distribution is recommended for discrete variables with a fixed conditional likelihood logit λ in Pawlowski's DSCM design[37]. For type-2 diabetes, we use neural networks to predict the conditional likelihood logit λ based on the causes of type-2 diabetes as in Eq 5.2.

$$d := f_D\left(\epsilon_D; [g, \hat{a}]\right) = \left(\text{ Gumbel-max }_{\lambda}([g, \hat{a}])\right)(\epsilon_D), \tag{5.2}$$

where ϵ_D denotes an exogenous noise for d and it follows Gumbel distribution $\epsilon_D \sim Gumbel(0, 1)$, Gumbel-max $_{\lambda}$ denotes a function $f(\epsilon_D; \lambda) = \underset{0 \le l \le 1}{\operatorname{argmax}} \left(\epsilon_D^l + \lambda_l\right)$, where λ is the predicted probabilities for participants with normal glucose metabolism(λ_0) and with type-2 diabetes(λ_1) by a fully-connected neural network $\lambda = NN_\theta([g, \hat{a}])$.

We use cross entropy as the loss function for this binary classification task as in Eq 5.3. We minimize the loss function to optimize the parameter λ.

$$L(\theta; d, [g, \hat{a}]) = -d \log\left(\epsilon_D^d + \lambda_d\right), \text{where } \lambda = NN_\theta([g, \hat{a}]) \tag{5.3}$$

Fundus images

For fundus images i, it is high-dimensional data. A shallow normalizing flow like the above is too simple to express it. However, a deep normalizing flow can be computationally expensive as all transformations are implemented in the same high data dimension. So we take the idea of an amortized and explicit structural equation according to Nick Pawlowski[37]. We separate the equation f_I into a shallow invertible component h_I and a deep non-invertible component g_I. And the corresponding exogenous noises are $\epsilon_I = (u_I, z_I)$. The structural causal model for fundus images is shown in Fig 5.3, and the deep structural equation is in Eq 5.4.

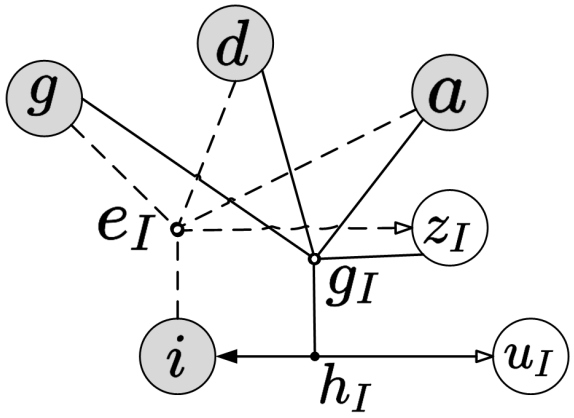


Figure 5.3: Amortized structural equation for fundus images

$$i := f_I\left(\epsilon_I; [g, \hat{a}, d]\right) = h_I\left(u_I; g_I\left(z_I; [g, \hat{a}, d]\right)\right), \quad p\left(\epsilon_I\right) = p\left(u_I\right) p\left(z_I\right) \tag{5.4}$$

In such a decomposition, we define a shallow invertible normalizing flow h_I. It maps a fundus image i into an exogenous noise of the same dimension u_I via a simple transformation. We also define a deep non-invertible decoder g_I mapping from a low-dimensional exogenous noise z_I and the parent variables of fundus images, $[g, \hat{a}, d]$ to the parameters of the transformation h_I. We expect the g_I to be designed complex enough to express the fundus image.

Note the conditional probability in this model, $p\left(i \mid [g, \hat{a}, d]\right)$ is intractable as the z_I is impossible to marginalize out. So for the low-dimensional exogenous noise z_I, we introduce a variational distribution $q\left(z_i \mid i, [g, \hat{a}, d]\right)$ to approximate its true distribution $p(z_I)$. We implement this variational inference with an encoder e_I. It can compress a fundus image and the parent variables of fundus images, $[g, \hat{a}, d]$, to the low-dimensional latent distribution over the possible z_I.

As a result, for the whole architecture consisting of e_I, g_I, and h_I, we can train it with variational inference by maximizing the ELBO, lower bound of $p\left(i \mid [g, \hat{a}, d]\right)$ in Eq 5.5.

$$\log p\left(i \mid [g, \hat{a}, d]\right) \geq \mathbb{E}_{Q\left(z_I \mid i, [g, \hat{a}, d]\right)}\left[\log p\left(i \mid z_I, [g, \hat{a}, d]\right)\right] - D_{\mathrm{KL}}\left[q\left(z_I \mid i, [g, \hat{a}, d]\right) \| p\left(z_I\right)\right], \tag{5.5}$$

$$p\left(i \mid z_I, [g, \hat{a}, d]\right) = p\left(u_I\right) \cdot \left|\det \nabla_{u_I} h_I\left(u_I; g_I\left(z_I, [g, \hat{a}, d]\right)\right)\right|^{-1}\Big|_{u_I = h_I^{-1}\left(i; g_I\left(z_I, [g, \hat{a}, d]\right)\right)} \tag{5.6}$$

Next, we introduce our design for the h_I, e_I, and g_I respectively. The shallow invertible normalizing flow h_I is designed as in Eq 5.7.

$$h_I\left(u_I; [s, \hat{a}, d]\right) = \left[\text{ Preprocessing } \circ \text{ ConditionalAffine }_{\theta}\left([g, \hat{a}, d]\right)\right]\left(u_I\right), \tag{5.7}$$

where Preprocessing denotes the inverted function of a fixed transformation function $f(x) = \operatorname{logit}\left(\alpha + (1-\alpha) \odot \frac{x}{256}\right), a = 0.05$. It maps the doubly bounded pixel value in the RGB channel to an unconstrained space. And the ConditionalAffine $_{\theta}\left([g, \hat{a}, d]\right)$ is an affine transformation that

maps the unconstrained space to the base distribution over u_I. We assume the base distribution to be independent multivariate Gaussian distribution. $u_I \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})), \boldsymbol{\mu} \in \mathbf{R}^k, \boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\sigma}^2) \in \mathbf{R}^k$, k is the pixel number of fundus image. In the experiment, we fix the log-variance for each variable as constant -5, $\log \sigma^2 = -5$. The parameter $\theta = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is predicted by a neural network decoder g_I with the input of the parent variables of fundus images, $[g, \hat{a}, d]$.

And we design the decoder g_I as in Fig 5.4 and the encoder e_I as in Fig 5.5.

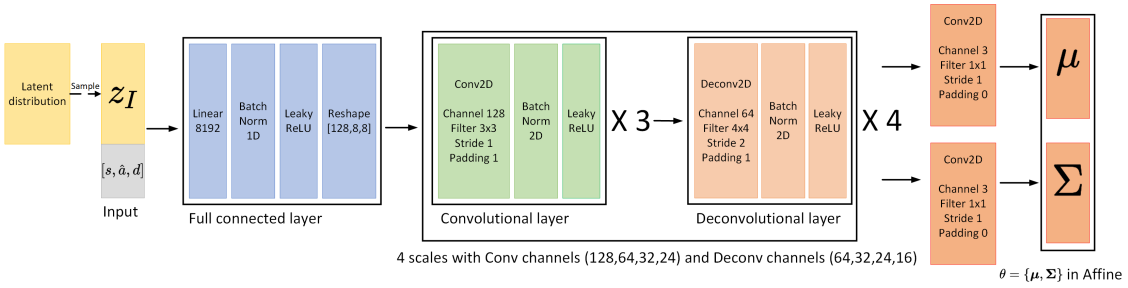


Figure 5.4: The structure of the decoder

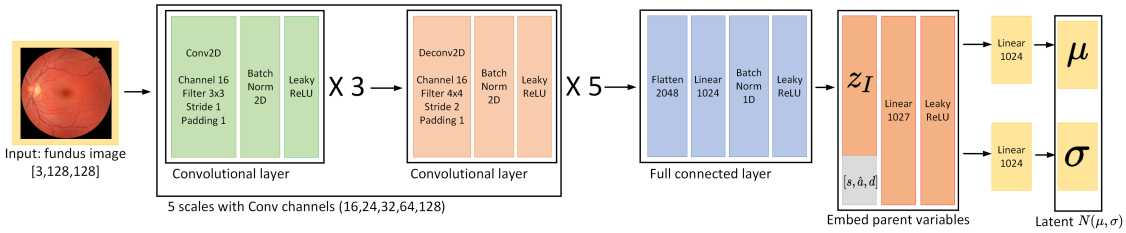


Figure 5.5: The structure of the encoder

5.2.3 Counterfactual inference using DSCM

Based on the trained deep structural causal model $\mathfrak{G}$, we implement the three steps for counterfactual inference in definition 3.2.6 with contrastive counterfactual conditions(see counterfactual inference on an example linear causal model in 3.2.4).

Abduction

Abduction is to predict the exogenous noise ϵ based on an individual observation $\mathbf{v}$. In our model, exogenous noises are independent according to definition 3.2.5, so the posterior can be inferred as $p_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{v}) = \prod_{k=1}^{K} p_{\mathfrak{G}}(\epsilon_k \mid v_k, \mathbf{Pa}(v_k))$. Namely, we can infer each exogenous noise separately for each structural equation based on its observed value and the value of its parent variables.

For variable age a, its structural equation is invertible so we can compute the exogenous noise ϵ_A deterministically with the inverted structural equation $\epsilon_A = f_A^{-1}(a)$.

For variable type-2 diabetes d, given that we observe $d = m$, we infer the exogenous noise by sampling from the posterior distribution $p(\boldsymbol{\epsilon}_D \mid d = m, \lambda)$. The sampling can be achieved by inverting the Gumbel-Max distribution according to the method by Chris J. Maddison [5] as Eq 5.8.

$$\begin{aligned} \epsilon_D^m &= G_m + \log \sum_l e^{\lambda_l} - \lambda_m, && G_m \sim \text{Gumbel}(0,1) \\ \epsilon_D^l &= -\log\left(e^{-G_l - \lambda_l} + e^{-\epsilon_D^m - \lambda_m}\right) - \lambda_l, && G_l \sim \text{Gumbel}(0,1), \quad \forall l \neq m \end{aligned} \tag{5.8}$$

For variable fundus image i, as the amortized and explicit structural equation is not deterministic, we use the encoder e_I to approximate the exogenous noise z_I. Besides, as the two exogenous noises, z_I and u_I, are not independent given the individual observation v, we approximate their joint distribution by replacing the independent distribution $p(u_I)$ by $p(u_I|z_I)$. Then the likelihood of exogenous noises can be approximated as in Eq 5.9.

$$\begin{aligned} p_{\mathfrak{G}}\left(\epsilon_I \mid i,[g,\hat{a},d]\right) &= p_{\mathfrak{G}}\left(z_I \mid i,[g,\hat{a},d]\right) p_{\mathfrak{G}}\left(u_I \mid z_I,i,[g,\hat{a},d]\right) \\ &\approx q\left(z_I \mid e_I\left(i;[g,\hat{a},d]\right)\right) \delta_{h_I^{-1}(i;g_I(z_I;[g,\hat{a},d]))}\left(u_I\right) \end{aligned} \tag{5.9}$$

Action

Action is to modify the deep structural equations $\mathbf{S}$ according to the desired intervention $\text{do}\left(v_k := \widetilde{v}_k\right)$. It results in a set of modified deep structural equations $\widetilde{\mathbf{S}}$. In our experiment, we implement five interventions respectively for an individual observation $\mathbf{v}$, which are different from its factual observational variables. For example, for an observational female participant at the age of 50 years old with type-2 diabetes $\mathbf{v}(g=0,d=1,a=50)$, we implement the following interventions: $\text{do}\left(g:=1\right)$, $\text{do}\left(d:=0\right)$, $\text{do}\left(a:=40\right)$, $\text{do}\left(a:=60\right)$, and $\text{do}\left(a:=80\right)$. We implement an intervention by replacing the structural equation with an assignment to the desired value.

Prediction

We have obtained the inferred exogenous noise $p_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{v})$ and the modified deep structural equations $\widetilde{\mathbf{S}}$ from abduction and action. They form a modified deep structural causal model $\widetilde{\mathfrak{G}}\left(\widetilde{\mathbf{S}}, p_{\mathfrak{G}}(\epsilon \mid \mathrm{v})\right)$. Based on the $\widetilde{\mathfrak{G}}$, we can perform the prediction by sampling from $\widetilde{\mathfrak{G}}$.

Note that we can preserve pixel-level details in the amortized and explicit structural equation for fundus images i. Take an example sample m via Monte Carlo Estimation as follows:

$$\begin{aligned} z_I^{(m)} &\sim q\left(z_I \mid e_I\left(i;[g,\hat{a},d]\right)\right) \\ u_I^{(m)} &= h_I^{-1}\left(i; g_I\left(z_I^{(m)};[g,\hat{a},d]\right)\right) \\ \widetilde{i}^{(m)} &= \widetilde{h}_I\left(u_I^{(m)}; \widetilde{g}_I\left(z_I^{(m)}; \widetilde{[g,\hat{a},d]}\right)\right) \end{aligned} \tag{5.10}$$

For the shallow invertible equation h_I, recall that we assume its base distribution follows independent Gaussian distribution. And we use the decoder to predict the parameters μ and σ for the element-wise affine transformation as $g_I\left(z_I;[g,\hat{a},d]\right) = \left(\mu\left(z_I;[g,\hat{a},d]\right), \sigma^2\left(z_I;[g,\hat{a},d]\right)\right)$. Then h_I can be parameterized as $h_I\left(u_I;\left(\mu,\sigma^2\right)\right) = \mu + \sigma \odot u_I$ (we omitted the fixed preprocessing transformation for brevity). Then the exogenous noise $u_I^{(m)}$ and the counterfactual image $\widetilde{i}^{(m)}$ in Eq 5.10 can be parameterized as follows:

$$\begin{aligned} u_I^{(m)} &= \left(i - \mu\left(z_I^{(m)};[g,\hat{a},d]\right)\right) \oslash \sigma\left(z_I^{(m)};[g,\hat{a},d]\right) \\ \widetilde{i}^{(m)} &= \mu\left(z_I^{(m)}; \widetilde{[g,\hat{a},d]}\right) + \sigma\left(z_I^{(m)}; \widetilde{[g,\hat{a},d]}\right) \odot u_I^{(m)} \end{aligned} \tag{5.11}$$

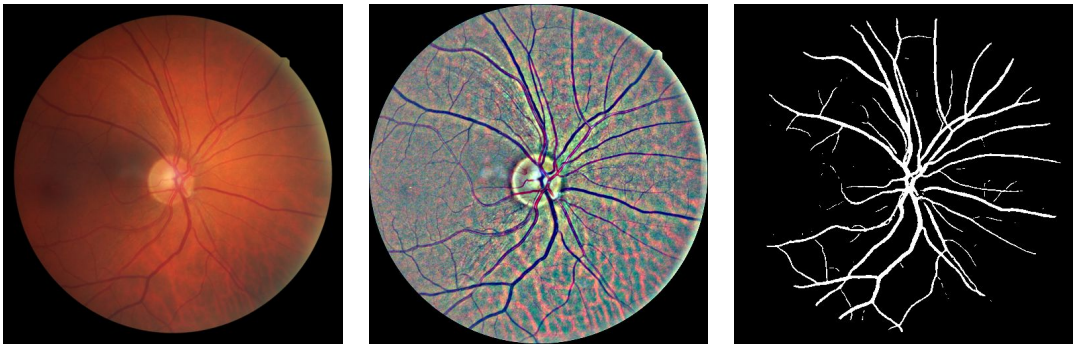
As we use the constant-variance $\log\sigma^2 = -5$, we can deduce that $\sigma\left(z_I^{(m)};[g,\hat{a},d]\right) = \sigma\left(z_I^{(m)}; \widetilde{[g,\hat{a},d]}\right)$. So the counterfactual image $\widetilde{i}^{(m)}$ can be deduced as Eq 5.12. In this way, the counterfactual image can preserve the pixel-level details.

$$\widetilde{i}^{(m)} = i + \left[\mu\left(z_I^{(m)}; \widetilde{[g,\hat{a},d]}\right) - \mu\left(z_I^{(m)};[g,\hat{a},d]\right)\right] \tag{5.12}$$

5.2.4 Image preprocessing

Image preprocessing allows us to remove unwanted distractions and improve specific qualities critical for the application we are working on. In our experiment, we adopt contrast normalization[13]

to enhance the anatomical structures. We also adopt a trained U-net model[43] to generate vessel mask. U-net is a successful biomedical segmentation model and can segment vessel out of fundus images with high accuracy. Note that the Pawlowski's DSCM design[37] is applied on gray images and does not adopt image preprocessing. We include image preprocessing to compare the expressiveness of the DSCM on different preprocessed fundus images. In our experiment, we use the original fundus images, the normalized fundus images, and the vessel mask of fundus images. We present a sample image for each in Fig 5.6.



(a) original fundus image (b) contrast normalized fundus image (c) vessel mask of fundus image

Figure 5.6: Sample fundus image after preprocessing

Contrast normalization

For contrast normalization, it enhances contrast by channel-wise local normalization. Concretely, given a pixel value $I(x, y)$ for a data point (x, y) in an image I and the neighbor N of the pixel. We normalize the pixel in the local neighbor by Eq 5.13.

$$\hat{I}(x, y) = \frac{I(x, y) - \mu_N(x, y)}{\sigma_N(x, y) + 20} \times 255 + 128, \tag{5.13}$$

where $\mu_N(x, y) = \frac{\sum_{(x,y)\in N} I(x,y)}{|N|}$ and $\sigma_N(x, y) = \sqrt{\frac{\sum_{(x,y)\in N} (I(x,y)-\mu_N(x,y))^2}{|N|}}$.

We set the local neighbor N as a square of side length 21 centered at the pixel, and we add 20 to $\sigma_N(x, y)$ as an adjustment. The normalization is implemented on each RGB channel.

Vessel Mask

For vessel mask, it segments the vascular structure out of the whole fundus image. In the experiment, we apply a trained deep learning model, U-Net, to segment the vessels automatically. U-net based approaches have succeeded in many biomedical image segmentation tasks. Furthermore, in the retina vessel segmentation task it reaches an accuracy of 0.98[16]. It encodes and decodes an input image to an output image of the same dimension. As in Fig 5.7, the symmetrical structure of encoder-decoder blocks extracts multi-level features, then it concatenates and propagates these features to the output via skip connection and a bottleneck layer. We adopt a trained U-net model of the same architecture on retina fundus images[1].

[1] https://github.com/nikhilroxtomar/Retina-Blood-Vessel-Segmentation-in-PyTorch

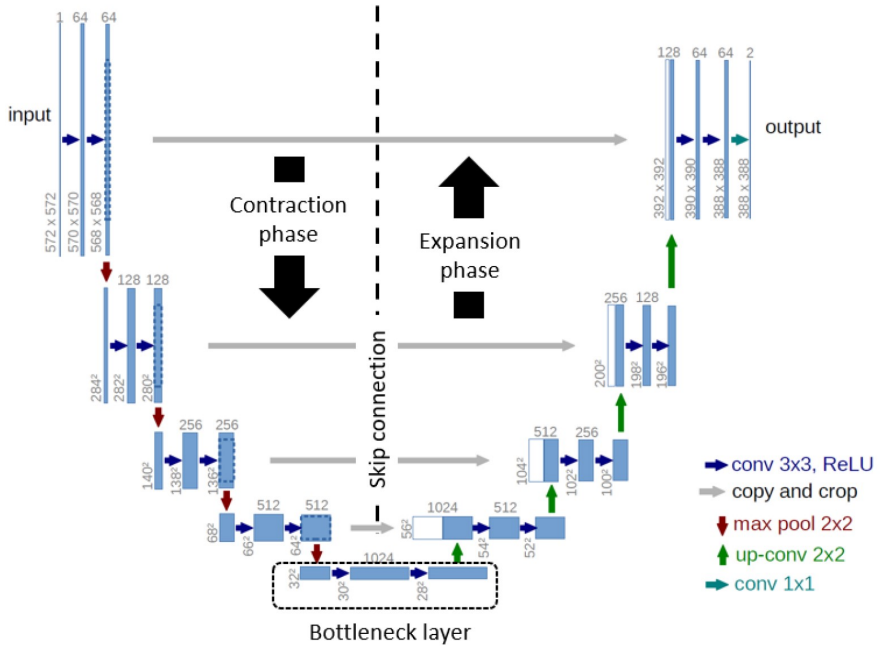


Figure 5.7: U-Net architecture. Image from [50]

Chapter 6

Experiment Result

In this chapter, we elaborate on the experiment results of our custom DSCM on the counterfactual inference task. We begin with the experiment setup. Then we evaluate the custom DSCM's image reconstruction and causal inference on the original fundus images. Moreover, we elaborate on the sensitivity analysis of the assumed causal DAG. After that, we compare different image preprocessing methods on fundus images and their impact on the custom DSCM's image reconstruction and counterfactual inference. Finally, we summarize the experiment results.

6.1 Experiment Setup

In this section, we introduce the experiment design, fundus image settings, training and evaluation parameter setting, and experiment infrastructure.

6.1.1 Experiment design

We split the experiments into three parts, DSCM on original fundus images, sensitivity analysis, and counterfactual inference on preprocessed fundus images.

DSCM on original fundus images

We train the custom DSCM on the training dataset for 1000 epochs with validating after each epoch on the validation dataset. Next, we reconstruct the fundus images. We encode the original fundus image and patient attributes to variationally infer the exogenous noise z_I. After that, we sample the exogenous noise u_I from the prior distribution. Then we reconstruct the original images based on the u_I and z_I. We evaluate the image reconstruction performance by comparing the reconstructed and original fundus images. The reconstructed images indicate the expressiveness of the encoded latent representation z_I. And the z_I is a core exogenous noise to infer in the abduction step in counterfactual inference. An exogenous noise with more image features encoded benefits the subsequent prediction step as more image features are utilized to make a counterfactual inference.

Moreover, we evaluate its causal inferences. For associational inference, it aims to model the joint distribution based on the statistical dependence in the observational data. We use the DSCM to sample the patient attributes from the prior distribution. After that, we evaluate the associational inference by comparing the joint distribution of the sampling patient attributes and the training dataset.

For interventional inference, it aims to model the causation by intervening on the causal DAG. We implement two experiments to evaluate the interventional inference. One is to compare the interventional distribution of the causes with different interventions on their effect. The other is to compare the interventional distribution of the effect with different interventions on their cause. With a valid interventional inference, the DSCM can block the causation from confounders, age

and gender to the cause, type-2 diabetes, so that the causation between type-2 diabetes and fundus images can avoid confounding.

For counterfactual inference, it aims to model the causation towards an individual observation under a counterfactual condition. As the counterfactual inferred fundus images are hypothetical retrospectives, we have no reference to evaluate them. Moreover, we sample the counterfactual inferred fundus images under contrastive counterfactual patient attributes as in section 5.2.3. Then we evaluate the relationships between the patient attributes and a fundus image by comparing the counterfactual inferred and the factual fundus images. Besides, we also evaluate the relationships quantitatively by linear regression of the mean pixel value per RGB channel on the age for the counterfactual inferred images.

Sensitivity analysis of the assumed causal DAG

We train DSCMs based on alternative causal DAGs and compare their performance of associational inference using the mean absolute error(MAE) between the original and reconstructed images and the ELBO(evidence lower bound)(see Eq 5.5).

DSCMs on preprocessed fundus images

We train DSCMs on the contrast-normalized fundus images and the vessel mask of the fundus images in the same way. Then we use the trained DSCMs to reconstruct the fundus images and sample counterfactual inferred fundus images. We want to increase the expressiveness of the DSCM on the fundus images by highlighting the anatomical structure of the fundus images. We evaluate their counterfactual inference performance by the expressiveness of the sampling fundus images.

6.1.2 Fundus image setting

We downsample the fundus images from size 512*512 to size 128*128 with bilinear interpolation to fit the computing performance of our experiment infrastructure. Note that even though the vessel mask is binary in size 512*512, it is continuous after downsampling. We use the bilinear interpolation to downsample the vessel mask as it preserves the most features from the original segmentation.

6.1.3 Training and evaluation parameter setting

For the parameter setting in training, we use Adam as an optimization strategy and take the batch size of 128. And we set a learning rate of 5^{-3} for the spline flow(see the architecture of the spline flow in the structural equation of age in section 5.2.2). For the structural equation for fundus images, we set a learning rate of 10^{-4} for the encoder e_I and the decoder g_I (see section 5.2.2 for the architecture of the encoder and the decoder). And we set the number of Monte Carlo Estimation for ELBO as 4. For the evaluation, we use the mean of 32 Monte Carlo samples to estimate each reconstruction and counterfactual inference. The training parameter setting refers to Pawlowski's deep structural causal model[37]. For the distribution plotting in the causal inference, we set the sample number as 10000.

6.1.4 Experiment infrastructure

All experiments are conducted on the High-Performance-Cluster (HPC) of the IMAG/e group, TU/e. The GPU is NVIDIA GeForce RTX 2080 Ti.

6.2 DSCM on Original Fundus Images

In this section, we elaborate on the experiment results of reconstruction and causal inferences using DSCM on original fundus images. Please see appendix A.1 for the training curves.

6.2.1 Reconstruction

We present the original and reconstructed fundus images in Fig 6.1.

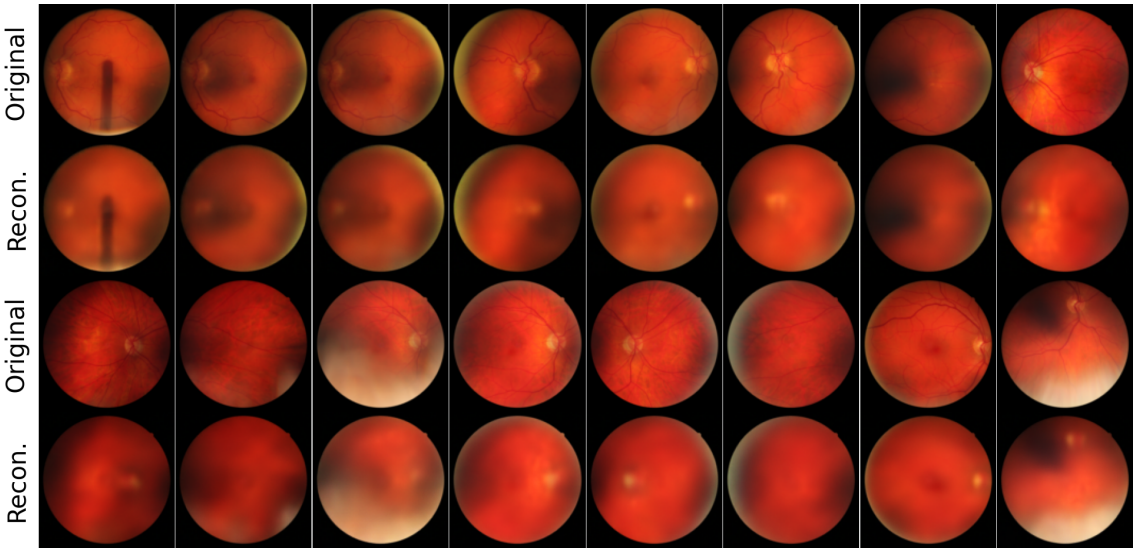


Figure 6.1: Reconstructed original fundus images

As shown above, the reconstructed images are generally blurry. This is a common issue for VAE as the ELBO tends to sacrifice the correct inference to overfit the training data under a limited capacity[60]. Moreover, it reconstructs color and limited anatomical structures like the optic disc and the macula. However, vessels and background texture are missing in the reconstructed fundus images. This implies that the encoded latent representation z_I has mainly captured the color feature of the fundus images.

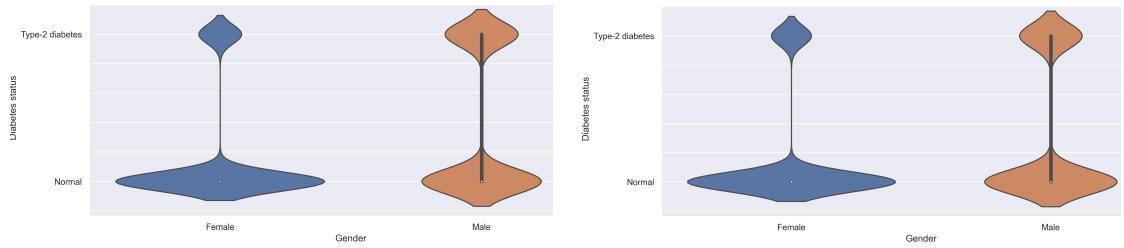
In conclusion, the DSCM trained on the original fundus images has expressiveness on the color in image reconstruction. This implies the encoded latent representation z_I has captured the color feature.

6.2.2 Associational inference

We use the DSCM to model the joint distributions by sampling exogenous noises from Gaussian distribution. We compare them with the joint distributions in the training data in Fig 6.2, Fig 6.3, Fig 6.4, Fig 6.5.

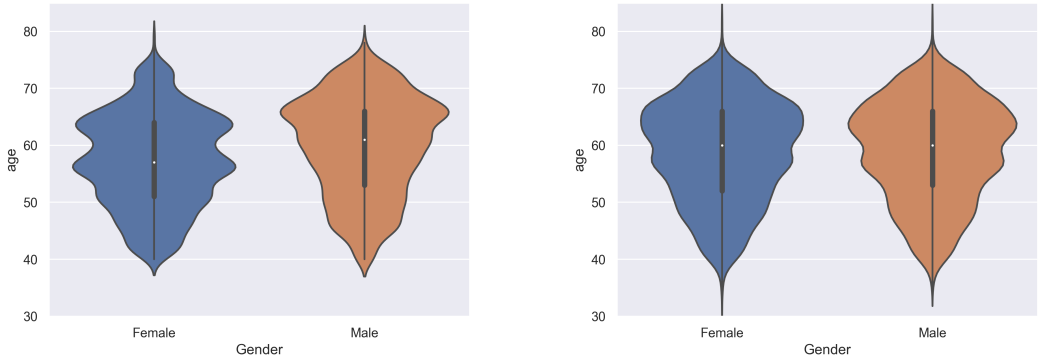
In Fig 6.2 and Fig 6.4, the joint distribution $p(d, g)$ and $p(a, d)$ are similar in the model and in the training dataset respectively. However, in Fig 6.3, the joint distribution $p(a, g)$ are different between the model and the training data. The main difference is $p(a|g = \text{female})$. And the joint distributions of age conditioned on different genders are the same in the model, $p(a|g = \text{female}) = p(a|g = \text{male})$. In Fig 6.5, the joint distribution $p(a, g, d)$ are different between the model and the training data. The main difference is $p(a|g = \text{female}, d)$.

In conclusion, the DSCM can model the joint distribution in the training dataset. However, it does not model the joint distribution of age conditioned on female $p(a|g = \text{female})$ in the training data. This is because age and gender are independent in our assumption and our model complies with this assumption. So in Fig 6.3(b), the distribution of age is independent of gender.



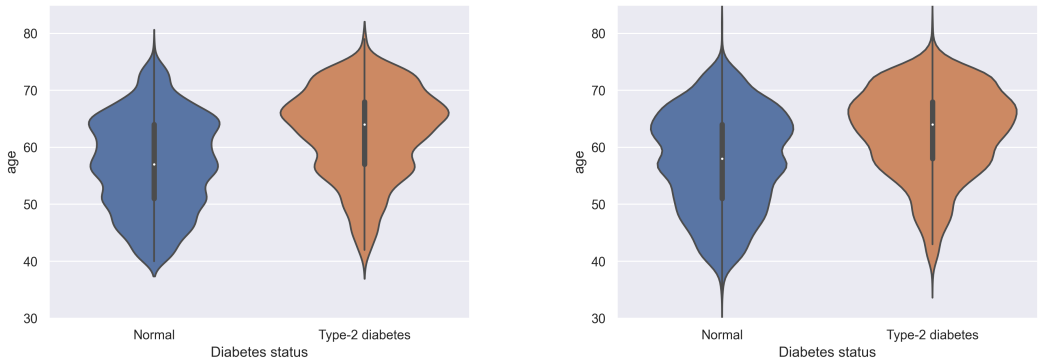
(a) Joint distribution $p(d, g)$ in training data (b) Joint distribution $p(d, g)$ modeled by the DSCM

Figure 6.2: Joint distribution of type-2 diabetes and gender $p(d, g)$ in training data and modeled by DSCM



(a) Joint distribution $p(a, g)$ in training data (b) Joint distribution $p(a, g)$ modeled by the DSCM

Figure 6.3: Joint distribution of age and gender $p(a, g)$ in training data and modeled by DSCM

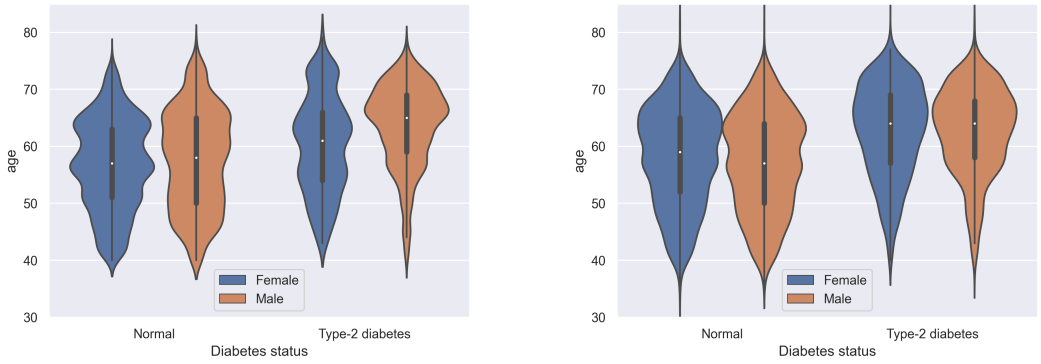


(a) Joint distribution $p(a, d)$ in training data (b) Joint distribution $p(a, d)$ modeled by the DSCM

Figure 6.4: Joint distribution of age and type-2 diabetes $p(a, d)$ in training data and modeled by DSCM

6.2.3 Interventional inference

In the assumed causal DAG in Fig 5.1, gender and age are the causes of type-2 diabetes. So intervention on type-2 diabetes blocks the causation from gender and age towards itself. And



(a) Joint distribution $p(a, g, d)$ in training data

(b) Joint distribution $p(a, g, d)$ modeled by the DSCM

Figure 6.5: Joint distribution of age, gender, and type-2 diabetes $p(a, g, d)$ in training data and modeled by DSCM

intervention on gender or sex does not affect the causation. We implement two intervention experiments to evaluate the causation after intervention and check whether it matches the expectation.

Firstly, we implement different interventions on the effect, type-2 diabetes. Moreover, we draw the interventional distribution of the causes, age $p(a|do(d))$ and gender $p(g|do(d))$ in Fig 6.6(a) and Fig 6.6(b). In Fig 6.6(a), the interventional distributions of age are the same, $p(a|do(d = \text{normal})) = p(a|do(d = \text{type-2 diabetes}))$. In Fig 6.6(b), the interventional distributions of gender are also the same, $p(g|do(d = \text{normal})) = p(g|do(d = \text{type-2 diabetes}))$. This is because the intervention on the DSCM blocks the causation from the age and gender towards the type-2 diabetes and makes them independent respectively. So the interventional distributions after intervention on the effect satisfy our expectation.

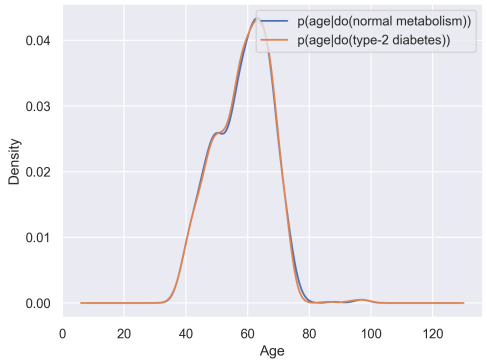
Secondly, we implement different interventions on the causes, age and gender. And we draw the interventional distribution of the effect type-2 diabetes, $p(d|do(a))$ and $p(d|do(g))$ in Fig 6.6(c) and Fig 6.6(d). In Fig 6.6(c), as we increase the desired value for the intervention on age, the ratio of type-2 diabetes increases accordingly. This shows that the DSCM preserves the causation from age toward type-2 diabetes after the intervention on age. In Fig 6.6(d), the ratio of type-2 diabetes after the intervention on the male is larger than on the female. This shows that the DSCM preserves the causation from gender toward type-2 diabetes after the intervention on gender. So the interventional distributions after intervention on the causes satisfy our expectations. Besides, we can conclude interventional inferences from the interventional distributions. For example, an older age and gender as male have positive causation on type-2 diabetes.

In conclusion, the custom DSCM can model causation by intervention and we can deduce interventional inferences from the interventional distribution.

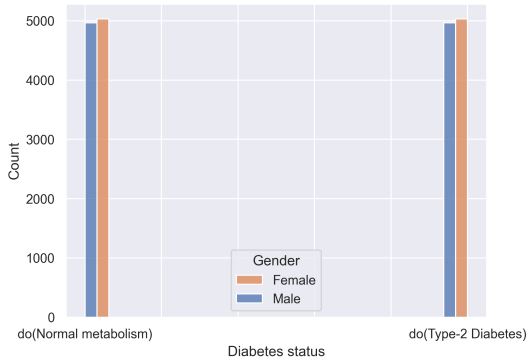
6.2.4 Counterfactual inference

We implement contrastive counterfactual inferences on the test dataset. For each observation, we make counterfactual inferences on the age of 40, 60, and 80 years old, the opposite gender, and the opposite diabetes status. Then we present the counterfactual inferred and the factual fundus images of two random observations in Fig 6.7.

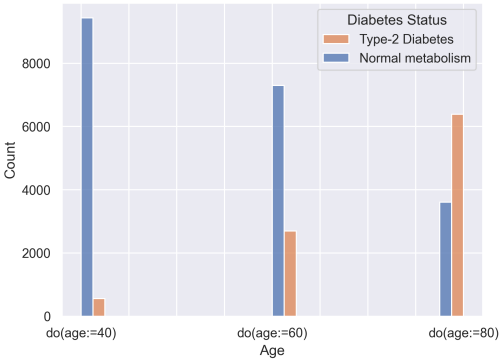
In Fig 6.7, we indicate the factual attributes of the observation in the head. And we make a counterfactual inference per column with the counterfactual condition at the column's title. The first row is the factual fundus images, and the second row is the counterfactual inferred fundus images. As the difference between the counterfactual inferred and the factual fundus images is small, for visualization purpose, we augment the counterfactual inferred fundus images



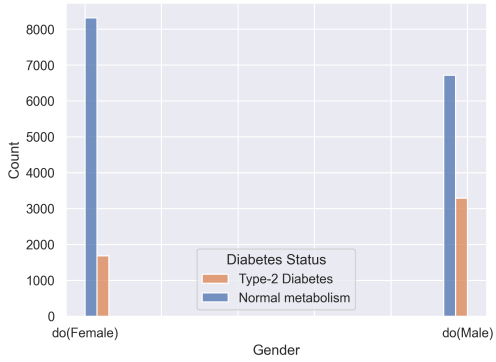
(a) Interventional distribution of age $p(a|do(d))$



(b) Interventional distribution of gender $p(g|do(d))$



(c) Interventional distribution of type-2 diabetes $p(d|do(a))$



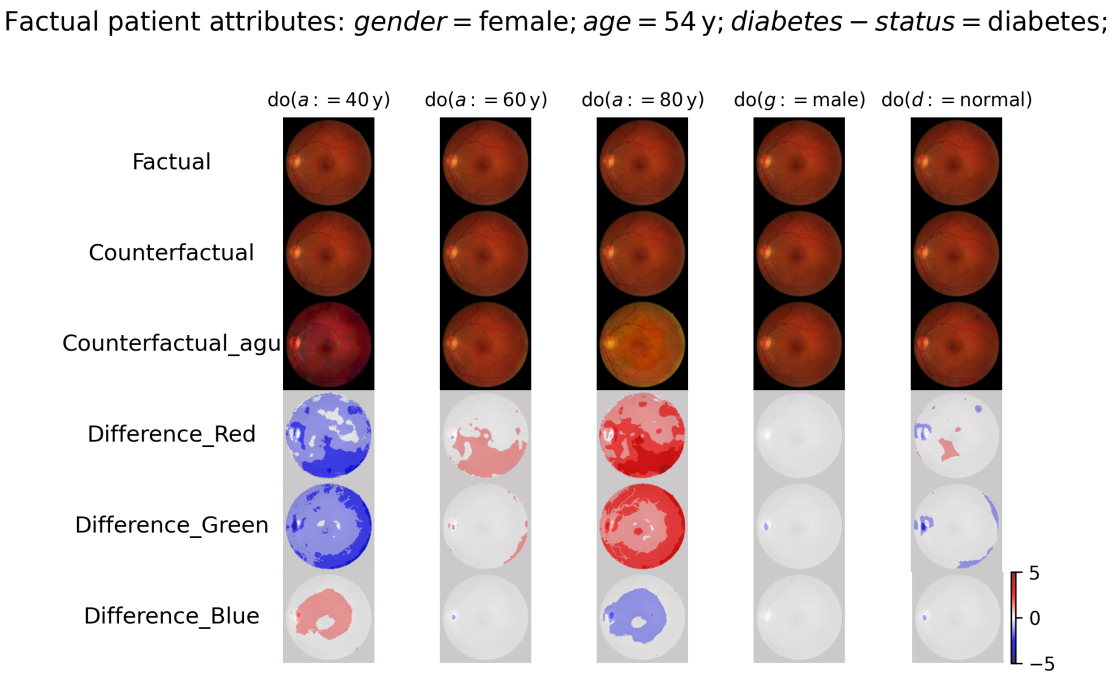
(d) Interventional distribution of type-2 diabetes $p(d|do(g))$

Figure 6.6: Interventional distribution using DSCM on original fundus images

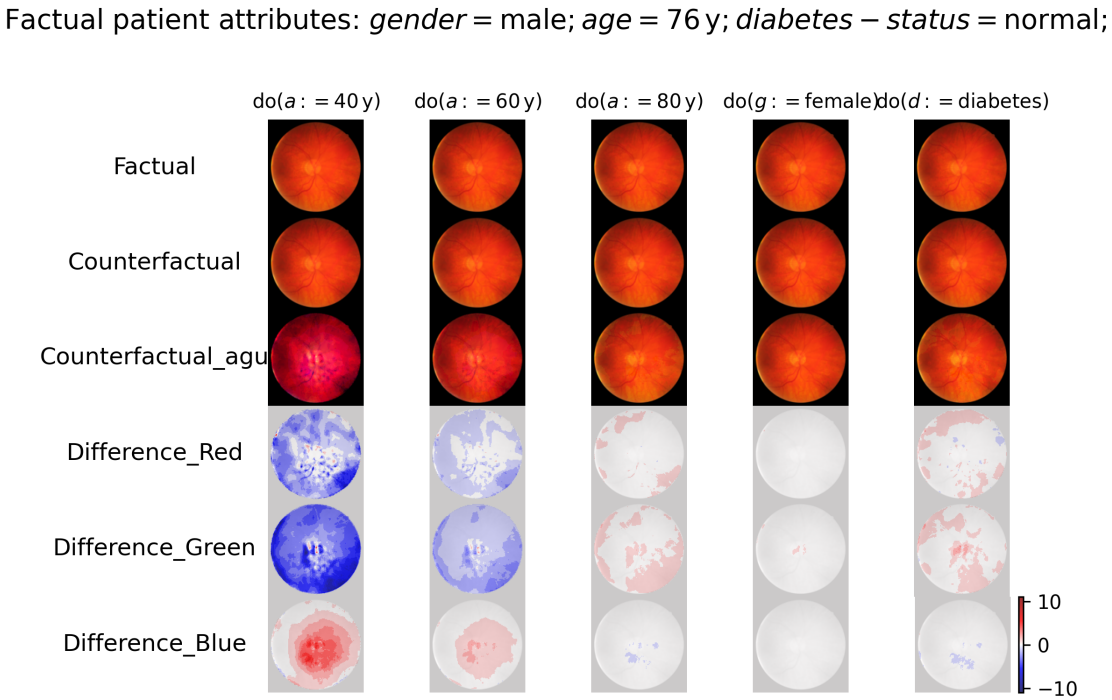
by multiplying the difference by 15 and adding it back to the factual fundus images. We draw the augmented counterfactual inferred fundus images in the third row. Next, we plot the difference per RGB channel in the last three rows.

In Fig 6.7(a), the observation is a 54-year-old female with type-2 diabetes. The counterfactual inferred fundus images have little difference from the factual fundus images. So we look into the augmented counterfactual fundus images. Compared with the factual fundus image, the counterfactual augmented inferred fundus image under 40 years old is more red as a result of the decreased pixel values in the red and green channels. And the counterfactual augmented inferred fundus image under 80 years old is more yellow as a result of increased pixel values in the red and green channels. The intervention on age of 60 years old has slight changes in the fundus images. This may be because the desired age, 60 years old, is close to the factual age, 54 years old. Moreover, the interventions on opposite gender and diabetes status have no informative changes in the generated fundus images.

In Fig 6.7(b), the observation is a 76-year-old male with normal metabolism. And the counterfactual inferred fundus images are also little different from the factual fundus images. Compared with the factual fundus image, the counterfactual augmented inferred fundus image on 60 years old is more red due to decreased pixel values in the red and green channels and increased pixel values in the blue channel. And the counterfactual augmented inferred fundus image on 40 years old is more red than the counterfactual augmented inferred fundus image on 60 years old. It has more decreasement in the red and green channels and more increasement in the blue channel.



(a) Counterfactual inference on a 54 years old female with type-2 diabetes

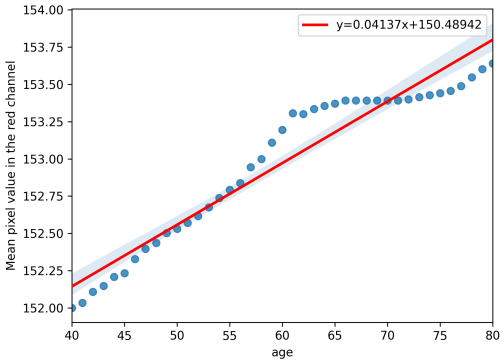


(b) Counterfactual inference on a 76 years old male with normal metabolism

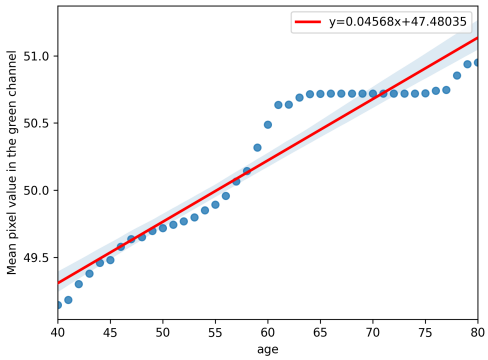
Figure 6.7: Counterfactual inferences using the DSCM on original fundus images

Besides, we find no informative changes in the counterfactual inference on gender and diabetes status.

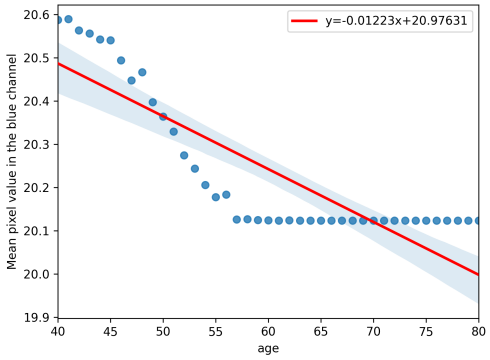
To quantify the change patterns on the color of the fundus images due to age, we make linear regression of the mean pixel value per RGB channel on age on an individual. Concretely, we make counterfactual inferences on an individual under the age range [40,80]. Then we compute the mean pixel value per RGB channel under different counterfactual ages. Next, we fit a linear regression of the mean pixel value per RGB channel on age and present them in Fig A.2. We also sample a population of 50 individuals and fit the same linear regression on the population in appendix A.2. It shows a similar change pattern as on an individual. From the linear regression, we can see that with the incresement of age, the pixel values in the red and green channels are increasing while the pixel value in the blue channel is decreasing. This makes the fundus images at a larger age more yellow as the color yellow is composed of red and green. This finding agrees with Bernhard's conclusion[12].



(a) Linear Regression of the mean pixel value in the red channel on age on an individual



(b) Linear Regression of the mean pixel value in the green channel on age on an individual



(c) Linear Regression of the mean pixel value in the blue channel on age on an individual

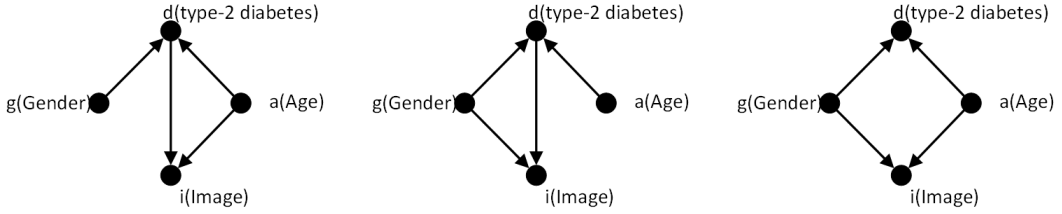
Figure 6.8: Linear Regression of the mean pixel value per RGB channel on age on an individual

In conclusion, the custom DSCM can model the counterfactual inference on an individual observation. The counterfactual inferred fundus images on age slightly differ from the factual fundus images. The counterfactual inferred fundus images at a larger age are more yellow than the factual fundus images, while the counterfactual inferred fundus images at a smaller age are more red than the factual fundus images. We quantify the change patterns with linear regression per RGB channel on age. And we deduce that aging causes the fundus images to be slightly more yellow due to increasement of pixel value in the red and green channels and decreasement of pixel value in the blue channel. However, the counterfactual inferred fundus images on gender and

type-2 diabetes have a bare difference from the factual fundus images.

6.3 Sensitivity Analysis

We implement a sensitivity analysis of the causal DAG assumption. Concretely, we construct DSCMs based on alternative causal DAG assumptions in Fig 6.9. After training, we evaluate their performance on the associational inference using the ELBO for $\log p(i, a, g, d)$ and $\log p(i|a, g, d)$ and MAE on the test dataset. MAE refers to the mean absolute error per pixel between the original images and the reconstructed images. The results are presented in Table 6.1.



(a) Alternative causal DAG 1 (b) Alternative causal DAG 2 (c) Alternative causal DAG 3

Figure 6.9: Alternative causal DAG assumptions

Model	Causes of fundus images	$\log p(i, a, g, d) \geq$	$\log p(i\|a, g, d) \geq$	MAE
Assumed causal DAG	age, gender, type-2 diabetes	-171867.71	-171862.76	2.59
Causal DAG 1	age, type-2 diabetes	-169511.53	-169506.58	**2.49**
Causal DAG 2	gender, type-2 diabetes	-173021.88	-173016.96	2.75
Causal DAG 3	age, gender	**-168872.18**	**-168867.23**	2.56

Table 6.1: Comparison of the model performance on associative inference, $\geq$ denotes the ELBO

In the alternative causal DAG 1, we count the gender out of the cause of the fundus images. And the DSCM trained on this assumption has an increased ELBO and the samllest MAE. In the alternative causal DAG 3, we count the type-2 diabetes out of the cause of the fundus images. Then the DSCM trained on this assumption has the largest ELBO and a slightly decreased MAE. Importantly, in the alternative causal DAG 2, after we count the age out of the cause of the fundus images, the DSCM trained on this assumption has the smallest ELBO and the largest MAE.

From the above results, we can conclude that the exclusion of age in the causes of fundus images degrades the model performance on associational inference, while the exclusion of gender or type-2 diabetes improves the model performance on associational inference. This is attributed to the limited expressiveness of the custom DSCM. Age has a significant causal effect on the color of the fundus images according to Bernhard's statistical modeling[12] , and the DSCM can reconstruct the color of the fundus images. Thus, the causation from age on the fundus images can be expressed by the DSCM. For type-2 diabetes and gender, people with type-2 diabetes have more tortuous retinal vessels according to M. B. Sasongko's retrospective study[47] and older people have a more elliptical optic disc and more reddish around it according to Takehiro's prospective observational study[58]. However, the detailed anatomical structures of the fundus images like this can not be well encoded into the latent representation z_I and reconstructed by the custom DSCM. Thus, the causation from gender and type-2 diabetes on the fundus images cannot be expressed by the DSCM.

In conclusion, the model performance on associational inference is affected by the design of causal DAG. Exclusion of age in the cause of fundus images degrades the model performance on associational inference while exclusion of gender or type-2 diabetes improves the model performance on associational inference. This is because the causation from age on the fundus images

is mainly on color and it can be expressed by the DSCM. While the causation from gender and type-2 diabetes on the fundus images is on the anatomical structure and it cannot be expressed by the DSCM.

6.4 DSCM on Processed Fundus Images

We train the DSCM on the contrast-normalized fundus images and the vessel mask of the fundus images. In this section, we elaborate on their image reconstruction and counterfactual inference performance.

6.4.1 Counterfactual inference on contrast-normalized fundus images

The contrast-normalized fundus images have enhanced vessels and background texture(see sample in Fig 5.6). We want the DSCM to capture these anatomical structures in reconstruction and apply it to counterfactual inference.

We present the reconstructed contrast-normalized fundus images in Fig 6.10. As shown below, the reconstructed images are blurry on the whole. It can preserve the yellow shadow, the optic disc, the truncated big vessels around the optic disc and partial background texture. We can conclude that contrast-normalization improves the expressiveness of the DSCM on the anatomical structure of the fundus images to a limited level. This implies that the encoded latent representation z_I has captured the partial anatomical stricture.

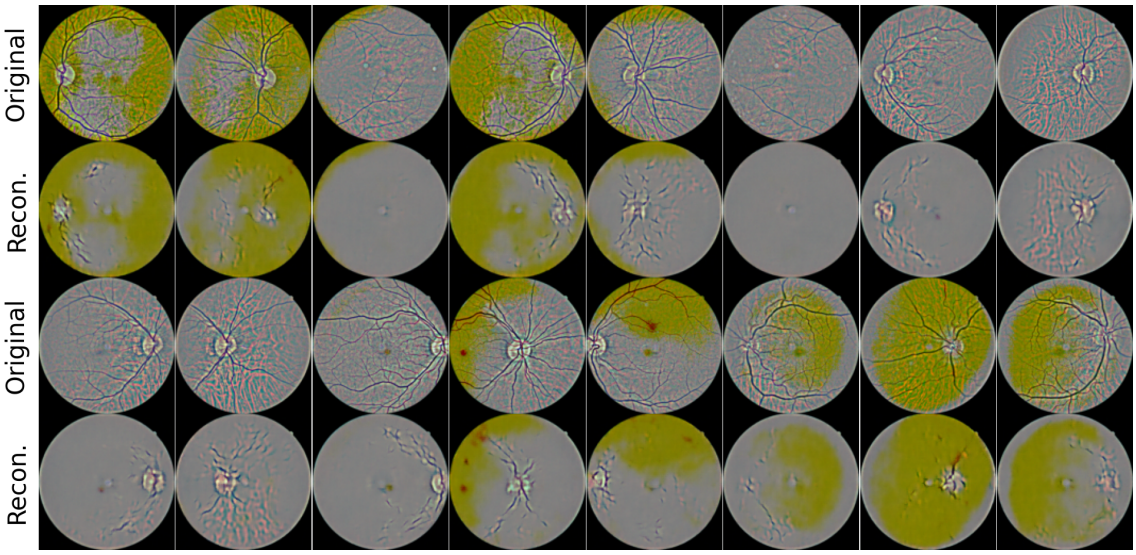
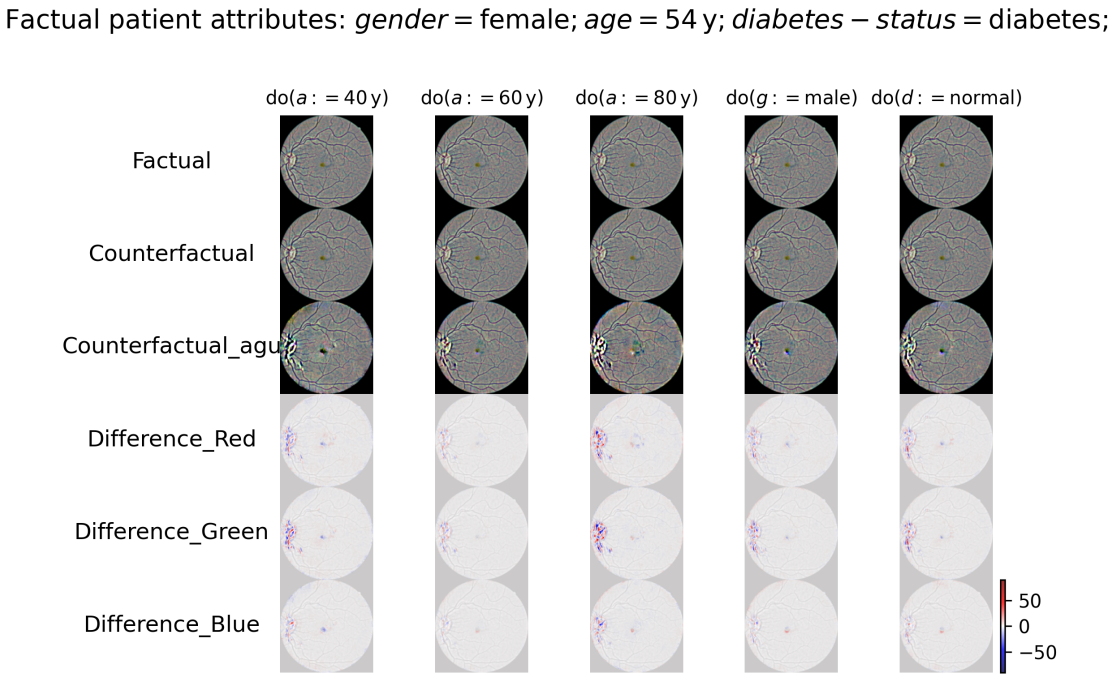
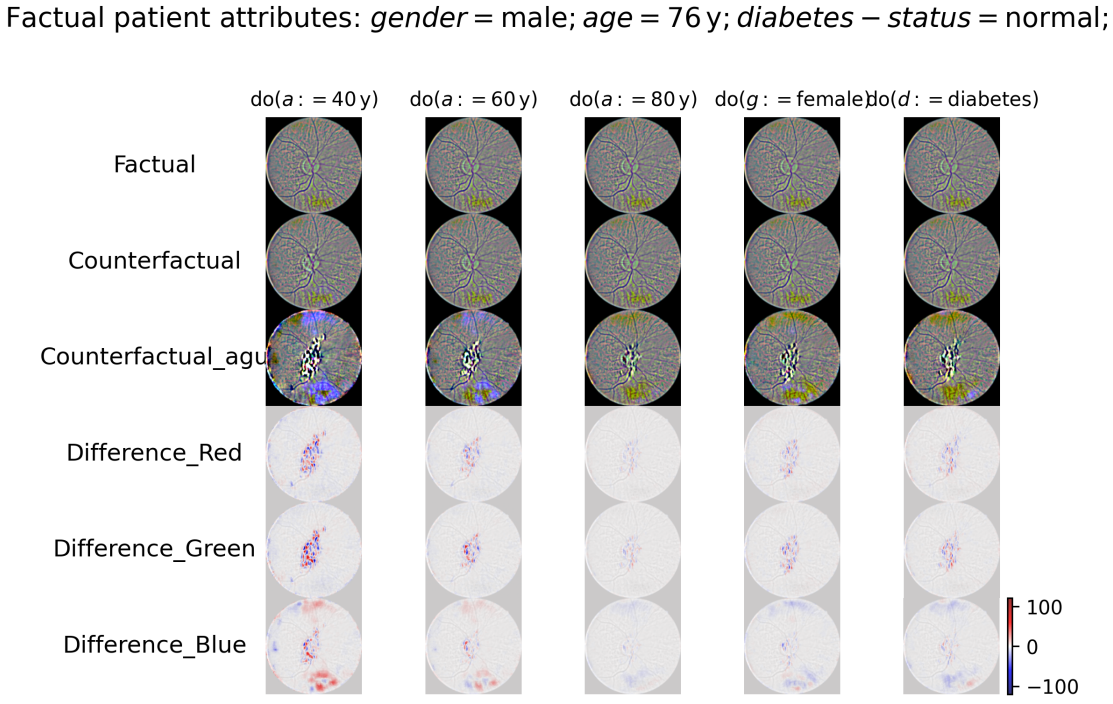


Figure 6.10: Reconstructed contrast-normalized fundus images

Next, we implement counterfactual inference on the same observations as in section 6.2.4 but with contrast-normalized fundus images. We present the counterfactual inferred contrast-normalized fundus images in Fig 6.11. In Fig 6.11(a), the counterfactual inferred fundus images with different interventions have slight changes in the same location. In the red and green channels, the changes are around the optic disc and along the vessels coming out from the optic disc. In the blue channel, the changes are located in the same place and the macula. The changes in the blue channel are smaller than in the red and green channels. And the changes are a cluster of randomly distributed increasement and decreasement in pixel values. In Fig 6.11(b), besides the similar changes in Fig 6.11(a), the counterfactual inference on the age of 40 and 60 increases the pixel value in the bottom right in the blue channel. Moreover, the counterfactual inferences on the female and type-2 diabetes decrease the pixel value in the bottom right in the blue channel.



(a) Counterfactual inference on a 54 years old female with type-2 diabetes



(b) Counterfactual inference on a 76 years old male with normal metabolism

Figure 6.11: Counterfactual inference using the DSCM on contrast-normalized fundus images

In conclusion, the DSCM trained on the contrast-normalized fundus images has improved expressiveness on the anatomical structure of the fundus images like the optic disc, the truncated big

vessels around the optic disc, and partial background texture in image reconstruction. This implies the encoded latent representation z_I has captured the partial anatomical stricture. However, we don't deduce informative change patterns in the counterfactual inferred images.

6.4.2 Counterfactual inference on vessel mask of fundus images

The vessel mask of fundus images only preserves the vessel shape out of the fundus images(see sample in Fig 5.6). We want the DSCM to reconstruct the vessel shape and apply it to the counterfactual inference.

We present the reconstructed vessel mask of fundus images in Fig 6.12. As shown, the reconstructed images are blurry on the vessel margins. And it preserves the main vessel trunk while losing small vessel branches. This implies the encoded latent representation z_I has encoded the vessel shape.

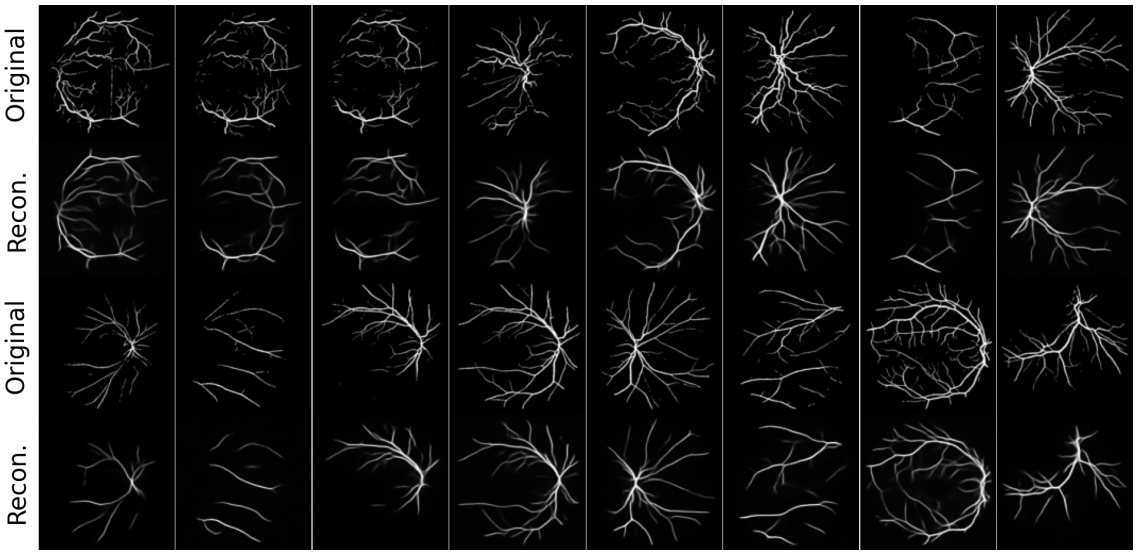
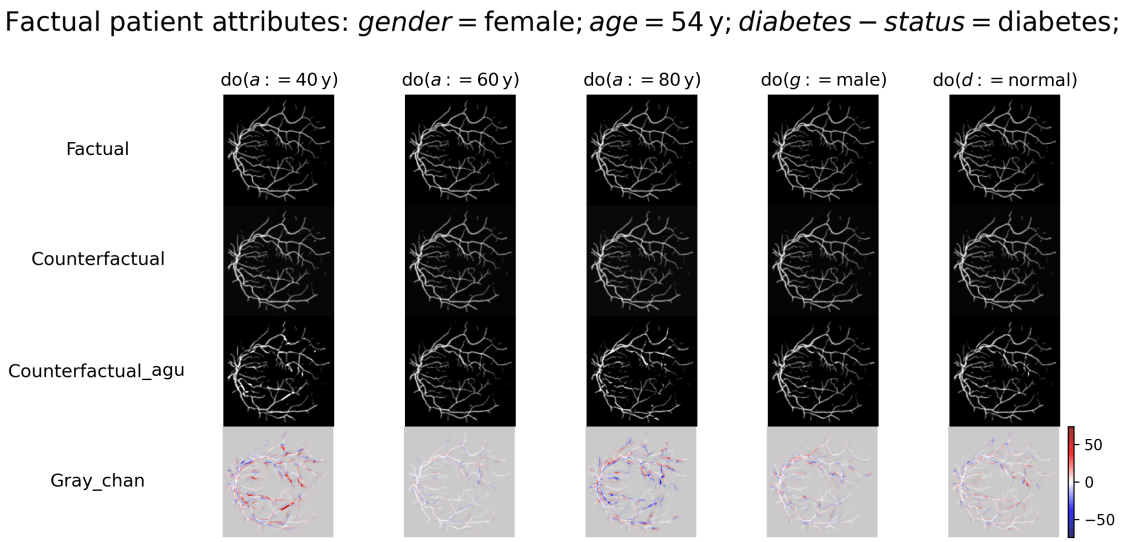


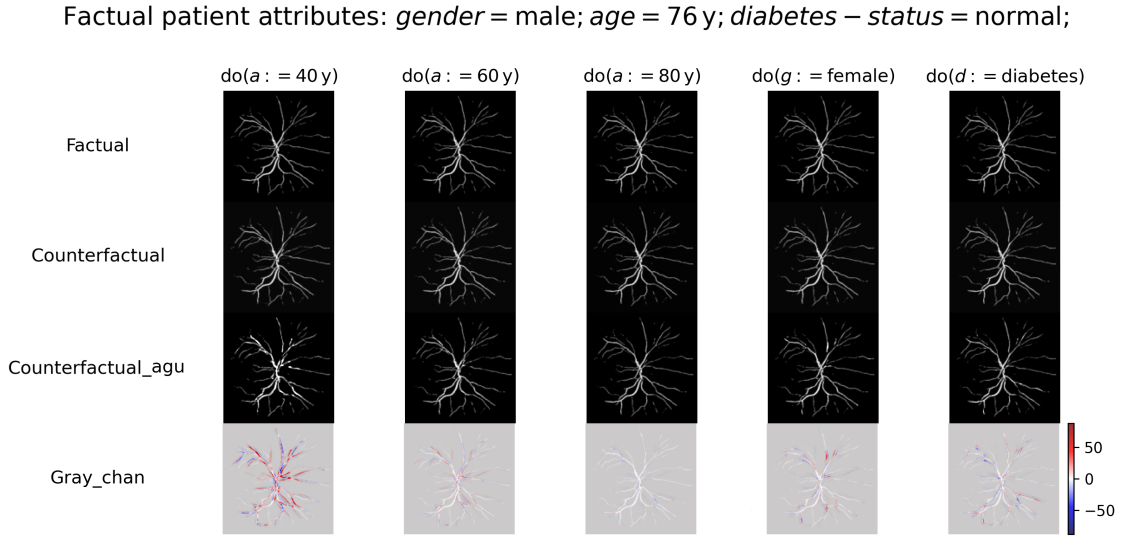
Figure 6.12: Reconstructed vessel mask of fundus images

Next, we implement counterfactual inference on the same observations as in section 6.2.4 but with the vessel mask of fundus images in Fig 6.13. In Fig 6.13, we augment the counterfactual inferred fundus images by multiplying the difference by 5 and then adding it back to the factual fundus images. And we plot the difference in the gray channel. In Fig 6.13(a), the difference between the factual and counterfactual inferred images is slight. In the augmented counterfactual inferred fundus images, the vessels are eroded strongly after the intervention on the age of 40 and 80 years old, and slightly after the intervention on the opposite gender and the diabetes status. The erosion is attributed to the discontinuous pixel changes along the vessels in the gray channel. In Fig 6.13(b), it follows the similar changes in Fig 6.13(a) except the erosion in the augmented counterfactual inferred fundus images is strongest when intervening on the age of 40 years old and less strong when intervening on the age of 60 years old.

In conclusion, the DSCM trained on the vessel mask of fundus images has expressiveness on the main vessel trunk of the fundus images in image reconstruction. This implies the encoded latent representation z_I has captured the partial vessel shape. However, we don't deduce informative change patterns in the counterfactual inferred images.



(a) Counterfactual inference on a 54 years old female with type-2 diabetes



(b) Counterfactual inference on a 76 years old male with normal metabolism

Figure 6.13: Counterfactual inference using the DSCM on vessel mask of fundus images

6.5 Summary of Experiment Results

In this section, we conclude and analyze the experiment results.

We train the custom DSCM on the Maastricht study and evaluate its image reconstruction and causal inference performance. We find that the custom DSCM has limited expressiveness in image reconstruction. The reconstructed images are blurry and preserve mainly color detail while losing anatomical structures. For the causal inference, we present that the custom DSCM can model the joint distribution in the training dataset and can model the causation by intervention. With the intervention, the custom DSCM can eliminate confounding from age and gender on the relation between type-2 diabetes and fundus images. Then we implement the contrastive counterfactual inference on fundus images. We find that the counterfactual inferred fundus images on a larger age are slightly more yellow and on a smaller age are slightly more red than the factual fundus

images. After quantifying the change patterns with linear regression per RGB channel on age, we conclude that aging causes the fundus images to be slightly more yellow.

Moreover, we implement a sensitivity analysis on the assumed causal DAG. We find that the exclusion of age from the causes of fundus images degrades the model performance on associational inference, while the exclusion of gender or type-2 diabetes improves the model performance on associational inference. We give our explanation for this. The causation from age on the fundus images is mainly on color and it can be expressed by the DSCM. While the causation from gender and type-2 diabetes on the fundus images is on the anatomical structure and it cannot be expressed by the DSCM.

To increase the expressiveness of the DSCM on the fundus images, we implement image preprocessing on the fundus images. They include contrast-normalized fundus images and the vessel mask of the fundus images. The DSCM on the contrast-normalized fundus images has improved expressiveness on the anatomical structure, like the optic disc, the truncated big vessels around the optic disc, and partial background texture in image reconstruction. Moreover, the DSCM on the vessel mask of the fundus images has expressiveness on the main vessel trunk in image reconstruction. Neither of them can sample informative and convincing counterfactual inferred fundus images. Thus, we conclude that image preprocessing improves the expressiveness of the DSCM on specific anatomical structure. However, it cannot improve the expressiveness of the counterfactual inference in our case. We think this is because the causation from type-2 diabetes and gender on the anatomical structure of the fundus images is small, and the DSCM model is not expressive enough to present detailed changes on it.

Chapter 7

Conclusions

In this section, we begin with a summary of the results from previous sections and an answer to the research question: **What is the causal relationships between the patient attributes and the fundus image appearance?**. After that, we discuss the reflection, limitations and future work.

7.1 Summary

The fundus images, a retina photograph, are a potential biomarker mapping to the patient attributes like age, gender, and type-2 diabetes. We attempt to explain the causal relationship between the patient attributes and the fundus images with two limitations mitigated. One is the confounding between the patient attributes, and the other one is the heterogeneity of fundus images. We construct a custom deep structural equation for each patient attribute based on the assumed causal DAG. Specially, we construct an amortized and explicit structural equation for the fundus images. We combine them to form a custom DSCM. Moreover, we intervene on the DSCM to avoid confounding. Besides, we make counterfactual inferences on the fundus images by introducing exogenous noises presenting the implicit individual variations. With contrastive counterfactual inferred fundus images, we discover the causal relationship between age and fundus images. The experiments on the Maastricht Study show that the custom DSCM can model the causation between age, gender, type-2 diabetes, and fundus images. Moreover, for the research question: **What is the causal relationships between patient attributes and the fundus image appearance?**, we give our finding: Aging causes the fundus images to be slightly more yellow.

7.2 Reflection

Our custom deep structural causal model(DSCM) models the causation between the patient attributes. It is promising to improve the transparency and convenience in the clinic. For example, doctors can quantify the causation from gender to type-2 diabetes by intervention as in Fig 6.6(d). Besides, doctors can infer whether the patient will have type-2 diabetes when he/she is older and take the corresponding strategy in advance. However, the custom DSCM has limited expressiveness on the fundus images. As a result, it can generate reasonable fundus images under counterfactual age but not under counterfactual gender and type-2 diabetes status. We think the bottleneck for the expressiveness of the DSCM lies in the latent encoded representation z_I. The z_I needs to encode more anatomical features, especially the vessels. After that, the DSCM can generate a reasonable counterfactual inference on gender and type-2 diabetes status.

Our finding that aging causes the fundus images to be slightly more yellow can be used to explain the black box behind neural networks. For example, Poplin[38] uses deep learning based on the fundus images to predict the patients' age accurately. The attention map indicates that the

important locations are scattered over the whole fundus images. It can be explained as follows. Aging causes the fundus images to have an increased pixel value in the red and green channels and a decreased pixel value in the blue channel over the entire fundus image. The neural network extracts these change patterns by convolutional layers and predicts age based on them.

We use contrast-normalization and vessel segmentation to highlight the anatomical structure of the fundus images and expect to encode more anatomical features into the latent encoded representation z_I. The custom DSCM does capture more anatomical structure in image reconstruction but fails in counterfactual inference. We think this is because the causation on the anatomical structure is tiny under counterfactual inference while the custom DSCM is not expressive enough to capture the detailed anatomical structure. We think highlighting the anatomical structure we desire to research on is a promising method to implement. By simplifying the counterfactual inference on the overall appearance of the fundus images to some specific anatomical structure, the encoded latent representation z_I can capture the desired anatomical structure.

7.3 Limitation

Bold assumption on the causal DAG

In our assumption, age and gender confound the relation between type-2 diabetes and fundus images. Moreover, we assume no unobserved confounder. This is a bold assumption. For example, hypertension is associated with type-2 diabetes and age[24, 51] but we don't include it. With a reasonable assumption on the causal DAG, the causation between patient attributes can be disentangled, and the counterfactual inferred images will be accurate.

Quality of the fundus images

The quality of the fundus images in the Maastricht Study can be improved. Some fundus images are occluded by a pillar, and some don't include anatomical structures like the optic disc and the macula. The quality of the fundus images would influence the density estimation of the DSCM and counterfactual inference.

Limited expressiveness of the custom DSCM

Our experiment shows that the expressiveness of the custom DSCM is limited to the color of the original fundus images. Moreover, for the contrast-normalized fundus images and vessel mask of the fundus images, the custom DSCM is not expressive enough to present the slight causation from type-2 diabetes and gender to the anatomical structure of the fundus images.

7.4 Future work

Assumption on the causal DAG

As our assumption on the causal DAG is bold, we expect to include more associated patient attributes and make a reasonable and convincing causal DAG based on the research finding on their relationships.

Quality of the fundus images

We expect to discard the fundus images of low quality. For the occluded fundus images, we can filter them out by rectangle detection. Moreover, for fundus images without the optic disc and macula, we can filter them out manually. And we expect to expand the data volume by merging other datasets, like RFMID[2].

[2]https://ieee-dataport.org/open-access/retinal-fundus-multi-disease-image-dataset-rfmid

Alternative causal models supporting counterfactual inference

As the custom DSCM has limited expressiveness on the fundus images, we expect to adopt an alternative causal model supporting counterfactual inference with better expressiveness, like the ImageCFGen(see section 4.2).

Bibliography

[1] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1):1–9, 2020. 1

[2] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988. 21

[3] Stephen A Burns, Ann E Elsner, and Thomas J Gast. Imaging the retinal vasculature. *Annual Review of Vision Science*, 7:129–153, 2021. 1

[4] BCK Choi and F Shi. Risk factors for diabetes mellitus by age and sex: results of the national population health survey. *Diabetologia*, 44(10):1221–1231, 2001. 7, 32

[5] Danny Tarlow Chris J. Maddison. Gumbel machinery. 35

[6] Gregory Cooper. An overview of the representation and discovery of causal relationships using bayesian networks. *Computation, causation, and discovery*, pages 4–62, 1999. 3

[7] Gianluca Coppola, Antonio Di Renzo, Lucia Ziccardi, Francesco Martelli, Antonello Fadda, Gianluca Manni, Piero Barboni, Francesco Pierelli, Alfredo A Sadun, and Vincenzo Parisi. Optical coherence tomography in alzheimer's disease: a meta-analysis. *PloS one*, 10(8):e0134750, 2015. 1

[8] Saloni Dash, Vineeth N Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 915–924, 2022. 28, 29

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 30

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 18

[11] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019. 20

[12] Bernhard M Ege, Ole K Hejlesen, Ole V Larsen, and Toke Bek. The relationship between age and colour content in fundus images. *Acta Ophthalmologica Scandinavica*, 80(5):485–489, 2002. 1, 46, 47

[13] Marco Foracchia, Enrico Grisan, and Alfredo Ruggeri. Luminosity and contrast normalization in retinal images. *Medical image analysis*, 9(3):179–190, 2005. 2, 36

[14] Edwin AM Gale and Kathleen M Gillespie. Diabetes and gender. *Diabetologia*, 44(1):3–15, 2001. 7, 32

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 24

[16] J Han, Y Wang, and H Gong. Fundus retinal vessels image segmentation method based on improved u-net. *IRBM*, 2022. 37

[17] Friso G Heslinga, Josien PW Pluim, AJHM Houben, Miranda T Schram, Ronald MA Henry, Coen DA Stehouwer, Marleen J Van Greevenbroek, Tos TJM Berendschot, and Mitko Veta. Direct classification of type 2 diabetes from retinal fundus images in a population-based sample from the maastricht study. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pages 383–388. SPIE, 2020. 1, 32

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 25

[19] P Insel and W Roth. Connect core concepts in health 13th brief edition. 2013. 7

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 21, 22, 27

[21] Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3964–3979, 2020. 19

[22] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951. 23

[23] Shinjini Kundu. Ai in medicine must be explainable. *Nature medicine*, 27(8):1328–1328, 2021. 1

[24] Rodrigo M Lago, Premranjan P Singh, and Richard W Nesto. Diabetes and hypertension. *Nature clinical practice Endocrinology & metabolism*, 3(10):667–667, 2007. 54

[25] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 18

[26] Yann LeCun et al. Generalization and network design strategies. *Connectionism in perspective*, 19(143-155):18, 1989. 22

[27] TJ MacGillivray, Emanuele Trucco, JR Cameron, Baljean Dhillon, JG Houston, and EJR Van Beek. Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. *The British journal of radiology*, 87(1040):20130832, 2014. 1, 7

[28] P Massin, JP Aubert, A Erginay, JC Bourovitch, A Benmehidi, G Audran, B Bernit, M Jamet, C Collet, M Laloi-Michelin, et al. Screening for diabetic retinopathy: the first telemedical approach in a primary care setting in france. *Diabetes & metabolism*, 30(5):451–457, 2004. 7

[29] Kevin McGeechan, Gerald Liew, Petra Macaskill, Les Irwig, Ronald Klein, Barbara EK Klein, Jie Jin Wang, Paul Mitchell, Johannes R Vingerling, Paulus TVM De Jong, et al. Prediction of incident stroke events based on retinal vessel caliber: a systematic review and individual-participant meta-analysis. *American journal of epidemiology*, 170(11):1323–1332, 2009. 1

[30] Franz H Messerli. Chocolate consumption, cognitive function, and nobel laureates. *N Engl J Med*, 367(16):1562–1564, 2012. 7, 8

[31] Sarah Müller, Lisa M Koch, Hendrik Lensch, and Philipp Berens. A generative model reveals the influence of patient attributes on fundus images. In *Medical Imaging with Deep Learning*, 2022. 1, 4

[32] Brady Neal. Introduction to causal inference. 13

[33] Eduardo Maria Normando, Benjamin Michael Davis, Lies De Groef, Shereen Nizari, Lisa A Turner, Nivedita Ravindran, Milena Pahlitzsch, Jonathan Brenton, Giulia Malaguarnera, Li Guo, et al. The retina as an early biomarker of neurodegeneration in a rotenone-induced model of parkinson's disease: evidence for a neuroprotective effect of rosiglitazone in the eye and brain. *Acta neuropathologica communications*, 4(1):1–15, 2016. 1

[34] 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. 33

[35] Abdulfatai B Olokoba, Olusegun A Obateru, and Lateefat B Olokoba. Type 2 diabetes mellitus: a review of current trends. *Oman medical journal*, 27(4):269, 2012. 7

[36] Niall Patton, Tariq Aslam, Thomas MacGillivray, Alison Pattie, Ian J Deary, and Baljean Dhillon. Retinal vascular image analysis as a potential screening tool for cerebrovascular disease: a rationale based on homology between cerebral and retinal microvasculatures. *Journal of anatomy*, 206(4):319–348, 2005. 1, 7

[37] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems*, 33:857–869, 2020. 1, 3, 27, 28, 33, 34, 37, 40

[38] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018. 1, 32, 53

[39] Stephen Powell. The book of why: The new science of cause and effect. pearl, judea, and dana mackenzie. 2018. hachette uk. *Journal of MultiDisciplinary Evaluation*, 14(31):47–54, 2008. 8

[40] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathologygan: Learning deep representations of cancer tissue. *arXiv preprint arXiv:1907.02644*, 2019. 1

[41] Sandeep Reddy. Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, 4(4):e214–e215, 2022. 1

[42] Julia M Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science*, 1(1):27–42, 2018. 7

[43] 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. 2, 37

[44] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985. 21

[45] PJ Saine, J Patrick, and ME Tyler. Ophthalmic photographers' society. *Fundus photography overview.[Cited 28 November 2013.] Available from: http://www. opsweb. org*, 2013. 1

[46] Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. *arXiv preprint arXiv:2202.10166*, 2022. 30

[47] MB Sasongko, TY Wong, TT Nguyen, CY Cheung, JE Shaw, and JJ Wang. Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. *Diabetologia*, 54(9):2409–2416, 2011. 47

[48] Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. 2022. 1

[49] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald Henry, and Coen DA Stehouwer. The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451, 2014. 31

[50] Narinder Singh Punn and Sonali Agarwal. Modality specific u-net variants for biomedical image segmentation: A survey. *arXiv e-prints*, pages arXiv–2107, 2021. 38

[51] Karri Suvila, Ville Langén, Susan Cheng, and Teemu J Niiranen. Age of hypertension onset: overview of research and how to apply in practice. *Current Hypertension Reports*, 22(9):1–8, 2020. 54

[52] Esteban G Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010. 18

[53] Brian L Trippe and Richard E Turner. Conditional density estimation with bayesian normalising flows. *arXiv preprint arXiv:1802.04908*, 2018. 18

[54] Noel S Weiss. *Clinical epidemiology: the study of the outcome of illness*, volume 36. Monographs in Epidemiology and, 2006. 3

[55] John P Wendland. The relationship of retinal and renal arteriolosclerosis in living patients with essential hypertension. *American Journal of Ophthalmology*, 35(12):1748–1752, 1952. 1

[56] Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019. 27

[57] Tien Yin Wong, Ronald Klein, A Richey Sharrett, Teri A Manolio, Larry D Hubbard, Emily K Marino, Lewis Kuller, Gregory Burke, Russell P Tracy, Joseph F Polak, et al. The prevalence and risk factors of retinal microvascular abnormalities in older persons: The cardiovascular health study. *Ophthalmology*, 110(4):658–666, 2003. 1

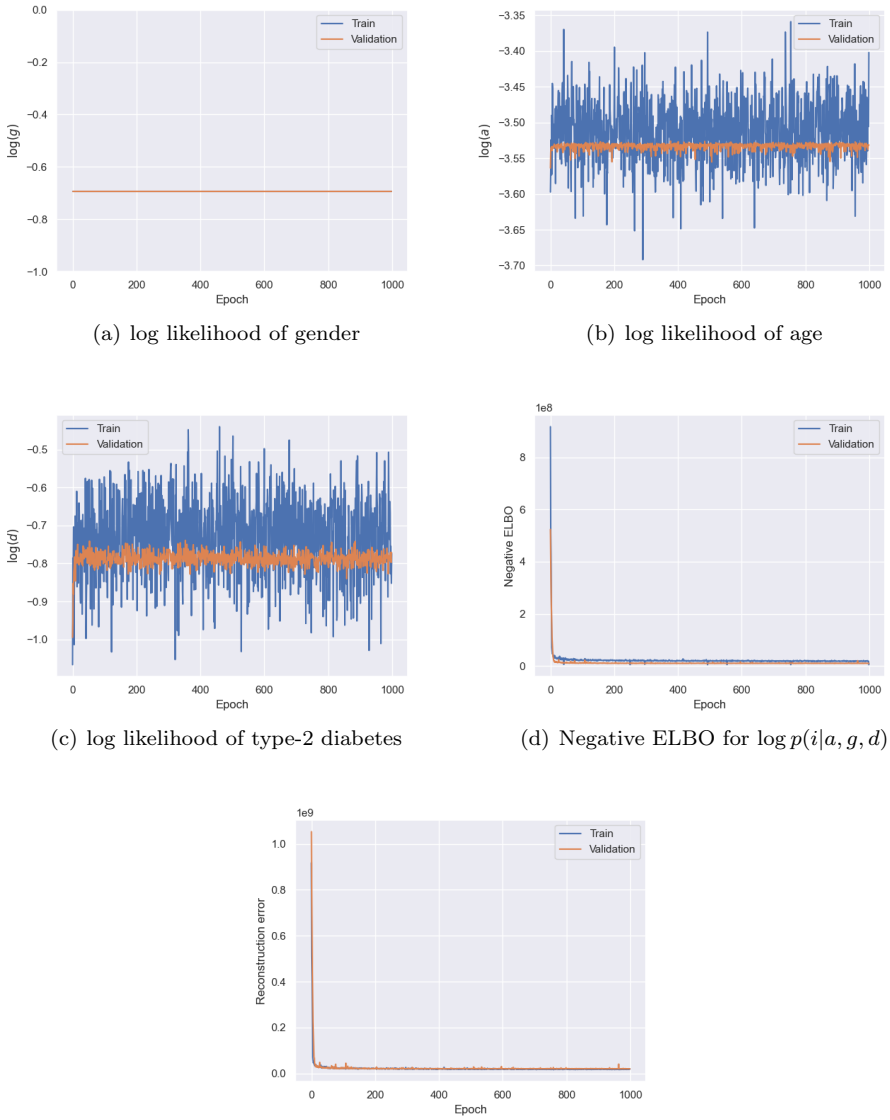
[58] Takehiro Yamashita, Ryo Asaoka, Hiroto Terasaki, Hiroshi Murata, Minoru Tanaka, Kumiko Nakao, and Taiji Sakamoto. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Translational Vision Science & Technology*, 9(2):4–4, 2020. 1, 47

[59] Kang Zhang, Xiaohong Liu, Jie Xu, Jin Yuan, Wenjia Cai, Ting Chen, Kai Wang, Yuanxu Gao, Sheng Nie, Xiaodong Xu, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nature Biomedical Engineering*, 5(6):533–545, 2021. 1

[60] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017. 41

Appendix A

A.1 Learning curves for DSCM on original fundus images



(a) log likelihood of gender

(b) log likelihood of age

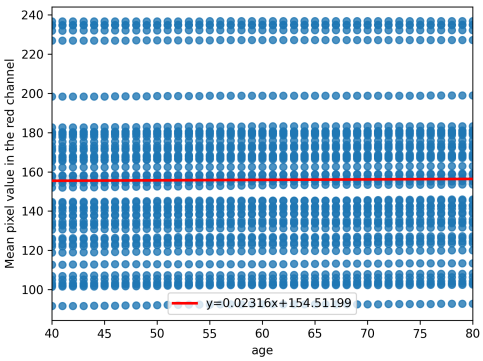
(c) log likelihood of type-2 diabetes

(d) Negative ELBO for $\log p(i|a, g, d)$

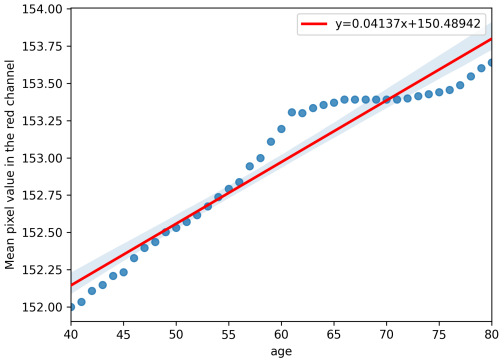
(e) Negative ELBO for $\log p(i, a, g, d)$

Figure A.1: learning curves for DSCM on the original fundus images

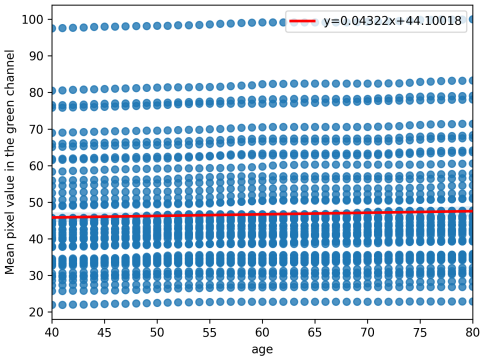
A.2 Linear Regression of mean pixel value per RGB channel on age on a counterfactual inferred population of 50 individuals



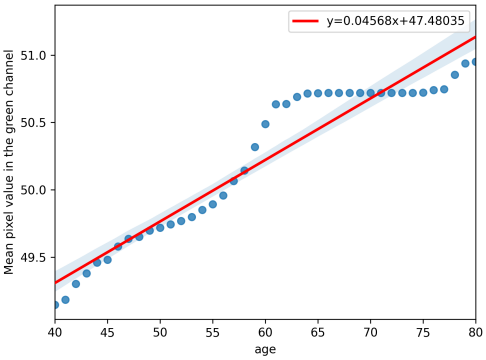
(a) Linear Regression of mean pixel value in the red channel on age on population



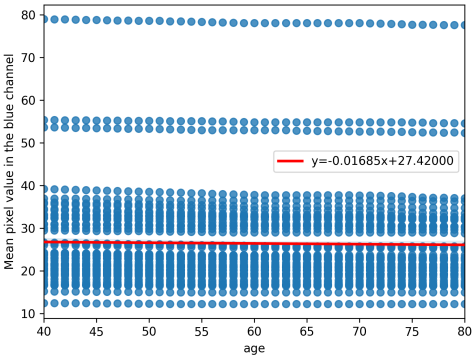
(b) Linear Regression of mean pixel value in the red channel on age on an individual



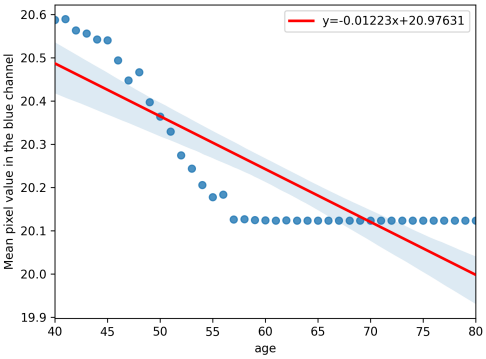
(c) Linear Regression of mean pixel value in the green channel on age on population



(d) Linear Regression of mean pixel value in the green channel on age on an indivudal



(e) Linear Regression of mean pixel value in the blue channel on age on population



(f) Linear Regression of mean pixel value in the blue channel on age on an indivudal

Figure A.2: Linear Regression of mean pixel value per RGB channel on age, (a)(c)(e) are on a sample population of 50 individuals, (b)(d)(f) are on an individual