Fair Image Generation Using β-IntactVAE

June 28, 2024

Student: Lars Chen
Matriculation Number: 0465415
Program: MSc Computational Neuroscience, TU-Berlin
Supervisors: Prof. Klaus-Robert Müller & Dr. Pengzhou (Abel) Wu

Abstract

Causal inference based on counterfactuals can help answer questions to hypothetical scenarios such as "How would this photo have looked if its subject were old instead of young?" Deep learning frameworks, such as the variational autoencoder (VAE), have often been used to predict alternative outcomes by intervening on a learned latent space with counterfactual treatments. However, generative models such as these are susceptible to learning spurious correlations in training data, leading to unfair representations, such as discriminatory image generation. Recently, machine learning has been integrated with causal inference to mitigate the effects of statistical biases. By ensuring that the distributions of control variables (e.g. demographic attributes) overlap significantly with treatment conditions, deep generative models have been shown to produce fairer representations. In this lab rotation, a novel VAE was constructed to model a prognostic score for imaging data which is sufficient for treatment effect estimation. By extending the framework of the β-Intact-VAE to a convolutional network, a generative prognostic model was trained on the CelebA dataset to predict counterfactual outcomes for facial images. Under certain conditions, the model was able to produce fair representations under limited overlap between control variables and treatment conditions.

1 Introduction

1.1 Motivation

In recent years, the field of image generation using machine learning (ML) has undergone significant advancements, driven by accessibility, cost-effectiveness, and broad applicability. From personal image editing to clinical research, high-fidelity image generation has become increasingly valuable (Ravi et al., 2019; Huang et al., 2022). One critical challenge, however, is the susceptibility of generative models to learning spurious correlations present in the training data, leading to unfair representations. Addressing these biases is crucial, particularly when such models are applied in sensitive areas like

healthcare and social media, where fairness and accuracy are particularly important (Rahmani et al., 2021).

1.2 Problem Statement

Some ML studies have recently relied on using causal information, such as structural causal models to control for confounding factors. However, in these cases, the causal graph must be given or assumed. While these techniques allow some level of control over biased learning, conditional image generation often samples labels independently, which ignores the interrelations among them (Pan et al., 2022; Huang et al., 2022). As a result, the final image produced might not accurately capture the relationships between label attributes. By ensuring significant overlap between the distributions of control variables (e.g., demographic attributes) and treatment conditions, deep generative models can produce fairer representations (D'Amour et al., 2020).

1.3 Research Objectives

Recently Wu and Fukumizu (2021) proposed β-Intact-VAE, a machine learning approach to model a prognostic score used for treatment effect estimation. β-Intact-VAE is a novel type of VAE that was used to learn conditionally balanced latent representations to predict counterfactual outcomes for low-dimensional data. This work aims to address the problem of biased image generation by extending the work done by Wu and Fukumizu (2021) to allow for counterfactual image generation. The primary objective is to produce fair representations under conditions of limited overlap between control variables and treatment conditions for imaging data. Specifically, this study seeks to:

- Extend the β-Intact-VAE using methods from computer vision, such as convolutions and residual layers
- Train our model on CelebA, a dataset containing facial images (Liu et al., 2018)
- Qualitatively evaluate the model's performance in generating counterfactual images that accurately reflect changes in specific attributes (e.g., age, gender) while maintaining fairness.

2 Methods

2.1 Causality and ML

Causal inference, referring to the process of estimating causal effects, has become a fundamental discipline of research (Pearl, 2009). Randomized controlled trials (RCTs) are considered the gold-standard protocols for estimating the effects of counterfactuals. By randomization, RCTs create two groups that are similar in all respects except for the intervention being applied. This procedure helps ensure that the comparison between groups is unbiased, allowing for a reliable estimate of the treatment effect. Thereby, predictions can be made for the treatment group if they had not received

the treatment, or vice versa, leading to the counterfactual scenario. While RCTs have been the main technique in investigating causal effects, the process of randomizing treatments is often unethical, not practical, or too expensive to perform(Rosenfeld et al., 2017). Thus, the inclusion of causal inference from observational data by using machine learning is important and has begun to garner attention. Although causal inference methods have been established in statistics-based fields such as epidemiology or econometrics, it has only recently been introduced into deep learning (Greenland et al., 1999; Schölkopf, 2022).

A fundamental problem in causal inference involves predicting answers to counterfactual questions that deal with hypothetical scenarios and aim at predicting outcomes of interventions done in the past. Counterfactuals, which may be posed as questions in the form "*What if X had happened instead of Y...*", often utilize generative models to predict such hypothetical outcomes for individual data points. However, inferring causal effects from observational data has its challenges. First, there may be *confounding*, caused by variables that are correlated with the independent variable and causally related to the dependent variable. Studies inferring causal effects from data typically control confounding by conditioning on the covariates. The more covariates are collected the more likely unconfoundedness holds. However, with the expanding covariate dimension, it is also more likely a strong imbalance exists between treatment and control (D'Amour et al., 2020). The other issue is systematic *imbalance* of the distributions of covariates, meaning the probability a subject belongs to a particular covariate group depends on whether it is in the treatment or control group (Rubin, 2005). Overcoming covariate imbalance is significant in improving the accuracy of generative image models. Real-world data is typically imbued with data imbalance due to societal biases and can cause under- or over-representation of a label. Classical studies tend to deal with *imbalance* using matching/re-weighting of data or learning a *balanced representation* Z which is independent of whether a data point belongs to treatment/control given its covariates (Stuart, 2010).

2.2 Causal Inference

The fundamental problem of causal inference is that, for a given experimental trial, we can only ever observe a single outcome $Y(t) \in \mathbb{R}^d$ per individual, where the factual treatment $T = t \in \{0, 1\}$ had occurred (Imbens and Rubin, 2015). $Y(t)$ is then seen as the variable that gives the *factual outcome* Y provided by the *factual assignment* $T = t$. Hence, only $Y(0)$ or $Y(1)$ may be observed given a single trial. However, $Y(t)$ is consistent with *counterfactual assignments* $T = t$ as well. Relevant covariates $X \in \{X_i \in \mathbb{R}^m\}$ are simultaneously observed per individual subjects. Realizations of an individual subject are distributed $p(X = \mathbf{x}, Y = \mathbf{y}, T = t)$, where upper-case signifies random variables and lower-case letters are their realizations.

This work mainly addresses causal inference under a limited overlap of covariates, a significant challenge due to the lack of data. Here, subjects with a certain covariate value almost belong exclusively to the treatment or control group. Some previous work mitigates the effects of limited overlap by trimming non-overlapping data points (Yang and Ding, 2018), or by weighting data points by over-

lap amount (Yao et al., 2018). Limited overlap occurs when propensity score $e(X) = P(T = 1, 0|X)$, which quantifies how likely an individual belongs to control or treatment given their covariates, is close to 1 or 0.

2.3 Related Work

VAEs learn low dimensional latent random variable Z which uses auto-encoding variational Bayes algorithm to infer intractable posterior $p(z|x)$ using VAEs (Kingma and Welling, 2022). Similar to auto-encoders, VAEs consist of an encoder network that compresses data into a low-dimensional latent space and a decoder that reconstructs the input data. However, with VAEs, the encoder and decoder output distributions over possible outputs (Zhang, 2018). The encoder network optimizes the variational distribution $q_\theta(z|x)$ and approximates $p(z|x)$. The decoder network represents the likelihood distribution $p_\phi(x|z)$, i.e. the generative model (Ganguly and Earp, 2021). VAEs are trained by maximizing the evidence lower bound (ELBO).

$$\begin{aligned} \log p(\mathbf{y}) &\geq \log p(\mathbf{y}) - D_{KL}(q(\mathbf{z}|\mathbf{y})||p(\mathbf{z}|\mathbf{y})) \\ ELBO &:= E_{\mathbf{z}\sim q}(\log p(\mathbf{y})) - D_{KL}(q(\mathbf{z}|\mathbf{y})||p(\mathbf{z})) \end{aligned} \tag{1}$$

The ELBO consists of a reconstruction error term, data log-likelihood, and the Kullback-Leibler divergence D_{KL} between $q_\theta(z|x)$ and $p(z)$, which is set to a parameterless standard normal distribution (Kingma and Welling, 2022), which regularizes the latent space.

Conditional VAEs (CVAE) build on the VAE by providing a conditioning variable, usually a class label, to the latent code passed to the decoder. During inference, the conditioning variable can be changed from the factual value to represent alternate classes (Sohn et al., 2015). Identifiable-VAE (iVAE) adds an auxiliary variable X, which contains observed data, to learn a parameterized factorized Gaussian prior distribution.

We focus on estimating treatment effects (TEs) on imaging outcomes Y, with binary labels T for treatment/control (i.e. non-treated), and other covariates X. Previously, Wu and Fukumizu (2021) proposed β-intact-variational autoencoder, a novel type of VAE that uses a prognostic score, to overcome limited overlapping covariates. Prognostic scores, a key concept for TE estimation, are a sufficient statistic of outcome predictors.

Counterfactuals predict hypothetical outcomes for individual subjects, hence β-Intact-VAE uses conditionally *balanced representation learning* (BRL) to map limited overlapping covariates X to an overlapping latent random variable Z, such that Z is independent of T given X. An important feature of β-Intact-VAE is that the latent variable is *identifiable*, meaning that the representation function and TEs are uniquely determined and expressed using the true observational distribution.

2.4 Data

To train the convolutional β-Intact-VAE, we used the large-scale CelebA dataset (Liu et al., 2018), a largely referenced dataset on computer vision works related to facial imaging. It contains more than 200K images from 10K different celebrity identities. CelebA data points contain 40 different binary labels.

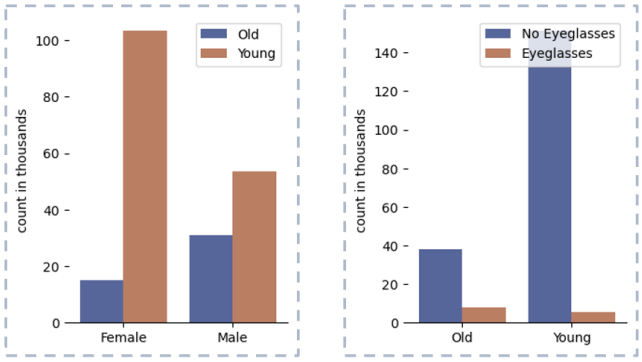

Figure 1. Imbalance in data labels.

CelebA contains images of famous individuals scraped off of the internet and labeled by a professional labeling company. This dataset was chosen since it contains data imbalances that reflect biases present in society. As per Figure 1, it can be seen that the imbalance between young and old subjects is more highly skewed for female subjects than males.

2.5 Model and Architecture

In this work, we build on the model used by Wu and Fukumizu (2021) to obtain a prognostic score for imaging data provided by the CelebA dataset. Our model combines the probabilistic graphical model of CVAE and iVAE. It is composed of three components: a decoder $p_\theta(\mathbf{y}|\mathbf{z},t)$, conditional prior $p_\lambda(\mathbf{z}|\mathbf{x},t)$, and an encoder $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},t)$. For inference, the loss function is given by the lower bound:

$$\log\, p(\mathbf{y}|\mathbf{x},t) \leq \mathbb{E}_{\mathbf{z}\sim q}(\log\, p_\theta(\mathbf{y}|\mathbf{z},t) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y},t)||p_\lambda(\mathbf{z}|\mathbf{x},t))) \tag{2}$$

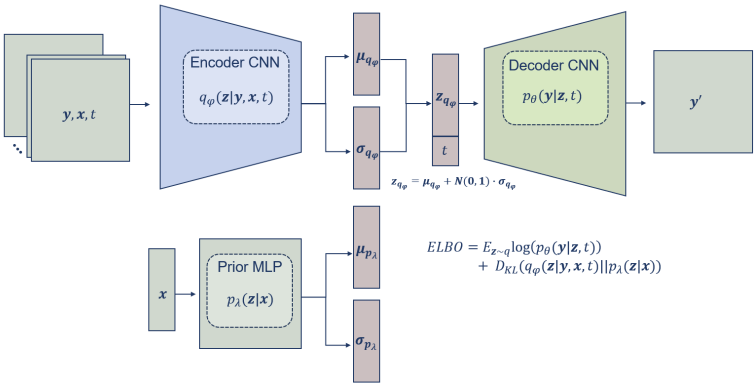

Figure 2. Convolutional β-Intact-VAE consists of variational distribution parameterized by an encoder CNN, prior distribution learned by an MLP, and the posterior parameterized by a decoder CNN.

Here, $p_\lambda(\mathbf{z}, \mathbf{x}, t)$ is a factorized Gaussian with dimensionality $dim(Z)$ and is parameterized by λ. We use a multi-layer perceptron for λ which consists of 3 hidden layers with a width of 256. The MLP output dimensionality is 256 which are split into vectors of 128 that describe the mean and log-variance of the conditional prior. The encoder $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}, t)$ is a convolutional neural network, shown in Figure 2, that consists of an initial convolutional layer with kernel size 3 and outputs 64 feature maps. This is followed by 5 residual blocks that consist of a BatchNorm, 2D convolution, BatchNorm, 2D convolution, and swish activation function as per Figure 3.

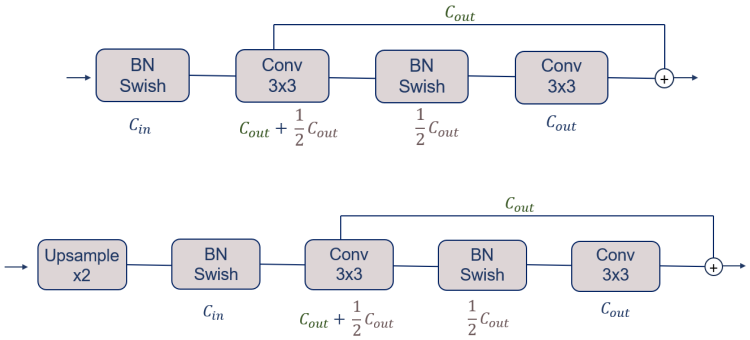

Figure 3. (top) The encoder CNN consists of 5 residual blocks successively compressing the data into smaller dimensions. (bottom) The decoder CNN consists of 5 upsampling residual blocks.

3 Results

Two counterfactual inference experiments are investigated where the observed image and observed covariates are passed to the encoder and any $t \in \mathcal{R}$ is passed to the encoder to sample a latent code

$\mathbf{z}$. $\mathbf{z}$ as well as t are passed to the decoder.

3.1 Gender Experiment

The first experiment investigated is counterfactual inference on gender. Typically, in other studies, $t = 0$ would refer to non-treatment and $t = 1$ to treatment applied. Here, we refer to t as the spectrum of the treatment feature. Hence, if the factual gender of a subject in an image is female $Y(t = 0)$, then a counterfactual outcome is said to be $Y(\hat{t})$ where $\hat{t} \neq 0$. Alternatively, if the factual gender is male, then the counterfactual outcome is said to be $Y(\hat{t})$ where $\hat{t} \neq 1$. The covariates learned, where $\mathbf{x_i} = 1, 0$ refer to positive and negative values of the following attributes: smiling, mouth open, beard, bald, pale, mustache, makeup, gray hair, bangs, high cheekbones, age.

Figure 4 shows examples of the model reconstructing images as either female or male. In the second column, both the reconstruction and counterfactual outcomes of a dark-haired man with blue/green eyes have the eye color changed to black. Similarly, the last example shows a blond-haired man with black eyes changing to more blue eyes. The reconstruction has darker blue eyes, whereas the young counterfactual outcome has light blue eyes. Additionally, reconstructing the image as female tends to make eyes lighter/bluer when the hair color is blond.

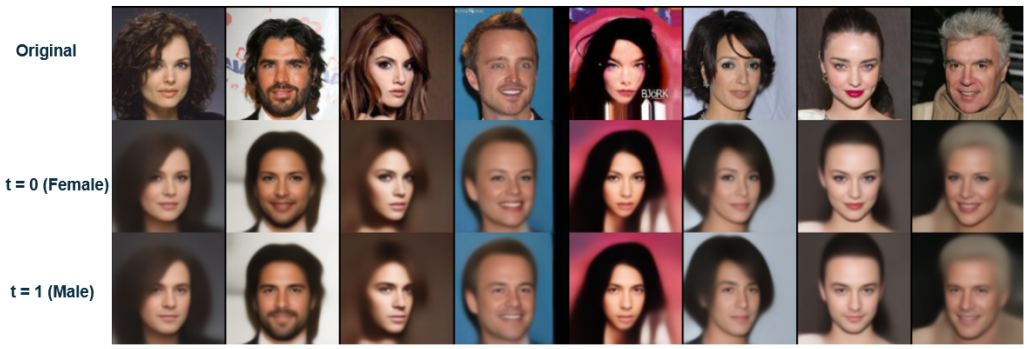

Figure 4. Select examples of (top) original image from the test set, (middle) reconstructing the image as female ($t = 0$), and (bottom) reconstructing the images as male ($t = 1$).

3.2 Age Experiment

The second experiment investigated is model age assignment. Here, if the factual age of a subject in an image is young $Y(t = 0)$, then a counterfactual outcome is said to be $Y(\hat{t})$ where $\hat{t} \neq 0$. Alternatively, if the factual age is old, then the counterfactual outcome is said to be $Y(\hat{t})$ where $\hat{t} \neq 1$. The covariates are the same as in the previous experiment, where age is switched. Examples are shown in 5. In this experiment, eye color is not preserved in the same examples as in the previous experiment.

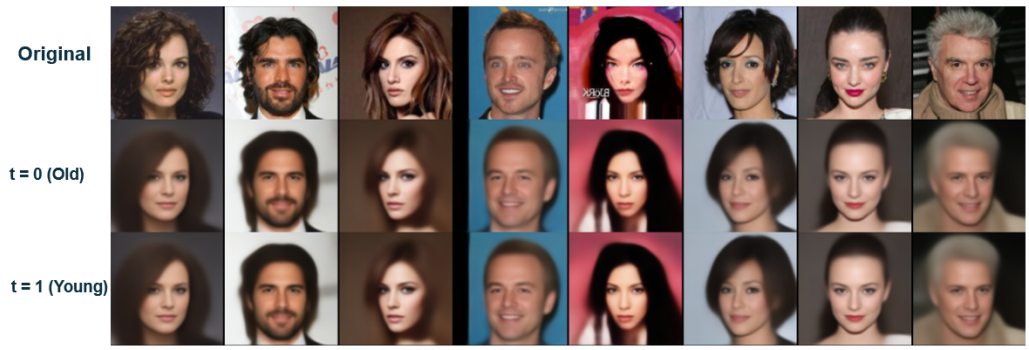

Figure 5. Select examples of (top) original image from test set, (middle) reconstructing the image as old ($t = 0$) and (bottom) reconstructing the images as young ($t = 1$).

3.3 Treatment Interpolation

Figure 6 shows interpolation over various treatment values for the gender and age experiments for two examples. When gender is varied, it can be observed that decreasing t creates a more female representation, and increasing t creates a more male representation. Covariate features in these two examples, such as age, paleness of skin, and smiling, appear to be preserved while t is in distribution. Some features are less preserved when t is set to further out-of-distribution, e.g. reduced smiling when t is increased. When age is varied, it can be observed that increasing t leads to younger representations. In this experiment, some covariate attributes are preserved in general such as gender, mouth openness, and paleness. Out-of-distribution t values that are too low, appear to reduce smiling.

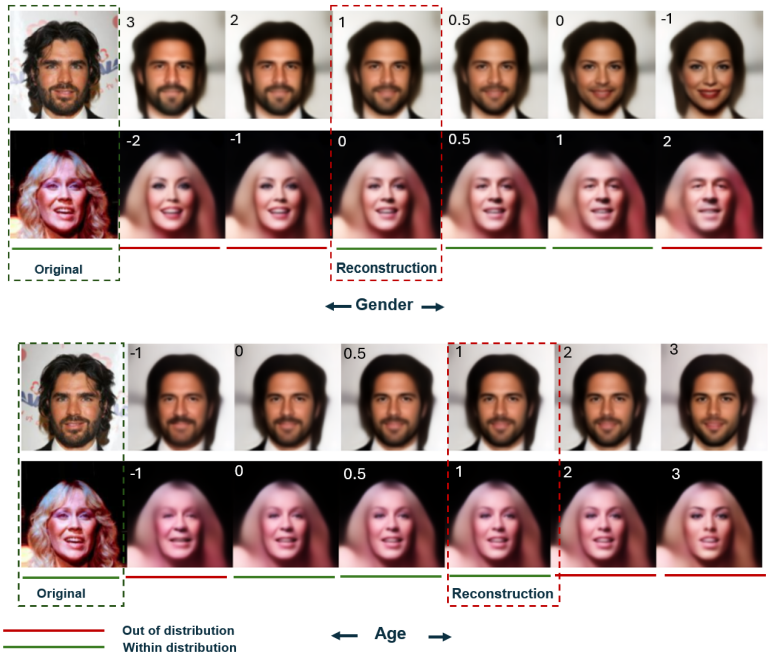

Figure 6. Original images are shown with the model output at different values of t. Reconstructions refer to model output where t corresponds to a null intervention. The leftmost images show the original data. Every image to the right shows a different t value, which includes in-distribution and out-of-distribution treatment values. (top) Two examples of counterfactual inference where gender is intervened. (bottom) The same two original images show counterfactual inference where age was intervened.

3.4 Latent Space Analysis

Figure 7 shows the latent distributions $p(\mathbf{z}|\mathbf{x}, \mathbf{y}, t)$ for two subjects, for two different values of t. It can be seen that changing T does not affect the latent distribution and that different subjects have unique latent codes.

4 Discussion

In this work, the β-Intact-VAE was adapted for image generation applications through the incorporation of convolutional layers. Our model demonstrated the capability to generate plausible counterfactuals for facial images in a few examples. The model was specifically tested on interventions involving age and gender, showing success in maintaining key attributes such as smile, mouth openness, and general facial structure, while modifying the target characteristics. Furthermore, the latent distribution does not change when the treatment value is adapted. This suggests that the convolutional

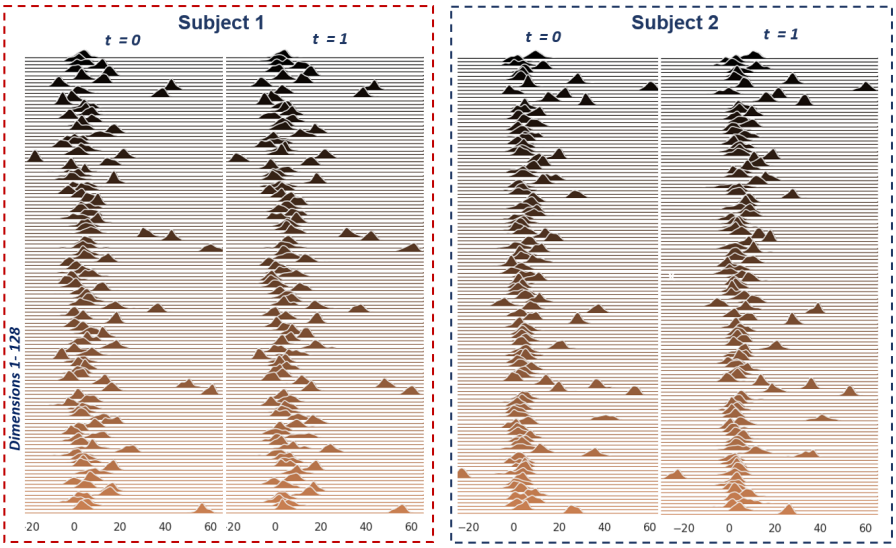

Figure 7. Factorized Gaussian plots showing probability density function of each latent dimension for two different subjects (Red/Blue) where the treatment is changed between male ($t = 1$) and female ($t = 0$)

β-Intact-VAE can learn a representation that is independent of the treatment variable, effectively separating the latent code from the treatment effect.

When the treatment variable was set to values outside the trained range [0,1], the model exhibited that it had learned spurious correlations. This was evident when male subjects were modified to appear younger, as the model inadvertently introduced more feminine features into the generated images. This suggests that while the model can generalize within the training range, it may lead to unintended attribute changes that were not part of the intended intervention.

Some limitations were noted in the model's performance, particularly in its ability to preserve eye color across counterfactual scenarios. This issue was more prominent in cases where the training data did not reflect the true diversity of eye and hair color combinations, highlighting the model's reliance on the biases present in the dataset. For example, when generating counterfactuals for individuals with dark hair, the model often defaulted to generating dark eyes, irrespective of the actual eye color in the original image. This indicates a spurious correlation learned from the training data, which predominantly associates dark hair with dark eyes and light hair with light eyes.

For the scope of this lab rotation, only a few examples were considered. To get a better sense of the model's capabilities for controlling spurious correlations, many more examples should be considered. Additionally, counterfactual results should be compared to those of a CVAE in addition to more state-of-the-art models.

Future work should explicitly control for additional covariates such as eye color to see if the model can eliminate identified spurious correlations. Additionally, since quantitative analysis was not in the

scope of this lab rotation project, further work should quantify and compare the performance of the current model with other state-of-the-art methods using metrics such as structural similarity index measure (SSIM) or Frechet Inception Distance (FID). Lastly, synthetic imaging datasets where the data-generating process is known should be considered.

Overall, the convolutional β-Intact-VAE represents a promising step in fair image generation, demonstrating the potential to produce unbiased counterfactual images in some instances. However, the findings also highlight the ongoing challenges in achieving truly fair and representative generative models, emphasizing the need for continued research and robust evaluation in this area.

References

A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning, 2020.

A. Ganguly and S. W. F. Earp. An introduction to variational inference, 2021.

S. Greenland, J. Pearl, and J. M. Robins. Causal diagrams for epidemiologic research. *Epidemiology*, pages 37–48, 1999.

S. Huang, Q. Li, J. Liao, L. Liu, and L. Li. An overview of controllable image synthesis: Current challenges and future trends. *Available at SSRN 4187269*, 2022.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2022.

Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Y. Pan, Z. Li, L. Zhang, and J. Tang. Causal inference with knowledge distilling and curriculum learning for unbiased vqa. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–23, 2022.

J. Pearl. *Causality*. Cambridge university press, 2009.

A. M. Rahmani, E. Yousefpoor, M. S. Yousefpoor, Z. Mehmood, A. Haider, M. Hosseinzadeh, and R. Ali Naqvi. Machine learning (ml) in medicine: Review, applications, and challenges. *Mathematics*, 9(22):2970, 2021.

D. Ravi, D. C. Alexander, N. P. Oxtoby, and A. D. N. Initiative. Degenerative adversarial neuroimage nets: generating images that mimic disease progression. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 164–172. Springer, 2019.

N. Rosenfeld, Y. Mansour, and E. Yom-Tov. Predicting counterfactuals from large historical data and small randomized trials. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 602–609, 2017.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

B. Schölkopf. *Causality for Machine Learning*, page 765–804. ACM, Feb. 2022. ISBN 9781450395861. doi: 10.1145/3501714.3501755. URL `http://dx.doi.org/10.1145/3501714.3501755`.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

P. Wu and K. Fukumizu. β-intact-vae: Identifying and estimating causal effects under limited overlap, 2021.

S. Yang and P. Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 2018.

L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. *Advances in neural information processing systems*, 31, 2018.

Y. Zhang. A better autoencoder for image: Convolutional autoencoder. In *ICONIP17-DCEC. Available online: http://users. cecs. anu. edu. au/Tom. Gedeon/conf/ABCs2018/paper/ABCs2018_paper_58. pdf (accessed on 23 March 2017)*, 2018.

5 Appendix

Table 1

Summary of training hyperparameters and settings.

Hyperparameter	Value
VAE learning rate	3e-4
Latent dimension	128
Training batch	16
Test batch	16
L2 decay	0
Fixed pixel log variance	1