Counterfactual Image Generation using GAN for Fairness

Koki Wataoka[†1,a)]   Takashi Matsubara[†2,b)]   Kuniaki Uehara[†3,c)]

Abstract: Computer vision systems have made significant improvements and been used in a variety of situations. For a practical use, we need to prevent the systems from making unfair decisions for certain individuals. In this sense, the systems have to eliminate the difference between decision makings on the real world and the counterfactual world where users would have different sensitive attributes (e.g., gender and race). In this study, we propose a framework for counterfactual image generation named Causality with Unobserved Variables using Generative Adversarial Networks (CUV-GAN). CUV-GAN can generate counterfactual images as the results of the intervention in the images' attributes and improve the fairness of an image classifier by being trained with generated images as data augmentation.

Keywords: generative adversarial networks, counterfactual fairness, causal inference, image manipulation

1. Introduction

Machine learning models have been used for a decision-making in many situations such as credit scoring [10], recidivism risk assessment [34], and recruitment [12]. When deploying a model, we have to ensure that the model does not discriminate against any individuals [13] [20] [23]. Hence, many studies have proposed various definitions of fairness in machine learning [3] [11] [17]. In this paper, we focus on a definition of fairness named counterfactual fairness [4] [25] [38]. Counterfactual fairness captures an intuition that a decision is fair towards an individual if it is the same in both the actual world and the counterfactual world where the individual belongs to a different demographic group.

Although you need counterfactual images to evaluate counterfactual fairness of your image classifier, the method of generating counterfactual images has not been studied much. Image manipulation techniques [6] [19] [27][33] [35] can edit attributes in images, but can not perform the intervention, which changes an attribute and other attributes according to the descendant nodes of the attribute by the strength of the causal relationships. We show the example of the difference between the manipulation and the intervention in Fig. 1.

Given the above, we propose a method of generating counterfactual images, named "Causality with Unobserved Variables using Generative Adversarial Networks" (CUV-GAN). CUV-GAN estimates structural equations for images and attributes and performs the intervention in the attributes of the images. We evaluate the quality of CUV-GAN in our experiments using a self-making dataset that has structural equations for the images and the attributes. In addition, we use the generated counterfactual images as data augmentation and confirm that CUV-GAN is effective for improving the fairness of classifiers.

2. Causality

2.1 Causal Model

Causal inference is a theoretical system for the elucidation of causal relationships among variables or events. We define a causal model according to Pearl [32].

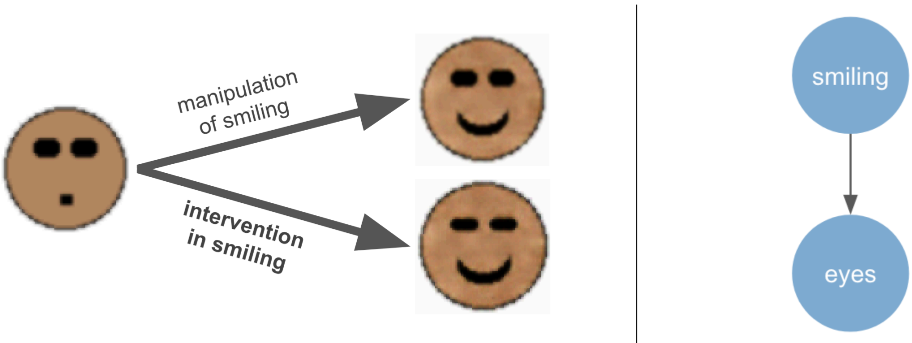
Definition 1 (Causal Model)**.** *A causal model is a triple (U, V, F) of sets such that*

- *U is a set of latent background variables, which are factors not caused by any variable in the set V of observable variables.*
- *F is a set of functions $\{f_1, ..., f_n\}$, one for each $V_i \in V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \backslash V_i$ and $U_{pa_i} \subseteq U$. Such equations are also known as structural equations.*

The aim of causal inference is to infer the structural equations and the distribution of latent background variables from observed variables.

Many studies [1] [24] [26] [36] [37] have investigated a basic causal model as follows: The causal relations of the ob-

[†1] Presently with Computational Science, System Informatics, Kobe University Graduate School
[†2] Presently with Graduate School of Engineering Science, Osaka University
[†3] Presently with Faculty of Business Administration, Osaka Gakuin University
[a)] wataoka@ai.cs.kobe-u.ac.jp
[b)] matsubara@sys.es.osaka-u.ac.jp
[c)] kuniaki.uehara@ogu.ac.jp



(a) Manipulation and intervention

(b) Causal diagram

Fig. 1 Comparison of the manipulation and the intervention. The manipulation of the smiling is to change only the smiling attribute in the image, and the intervention in the smiling is to change the smiling and eyes attributes in the image. (a) The left image is the original image. The upper right image is the result of the manipulation of the smiling by [27]. The lower right image is the result of the intervention in the smiling by our method. (b) The figure is a causal diagram of this example illustrating that the smiling attribute affects the eyes attribute.

served variables are graphically represented by a directed acyclic graph (DAG), that is, when the observed variables are in the causal order, no later variable determines any earlier variable in the DAG. Further, the functional relations of the observed variables are linear. We thus deprive structural equations of the basic causal model as follows:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^{d_y}$ denotes the observed variable, $\boldsymbol{B} \in \mathbb{R}^{d_y \times d_y}$ denotes the causal matrix collecting the connection strengths b_{ij} between y_i and y_j, and $\boldsymbol{e} \in \mathbb{R}^{d_y}$ denotes the exogenous variable, which is not considered as attributes but affects y. Here, the elements of the exogenous variable are assumed to be independent of each other. The basic causal model (1) is known as a linear non-Gaussian acyclic model, abbreviated as LiNGAM.

Furthermore, Hoyer *et al.* [18] extended the causal model (1) and formulated a linear acyclic structural equation model with latent confounders, named "Latent variable Linear Non-Gaussian Acyclic Model" (LvLiNGAM). Structural equations of LvLiNGAM is given by

$$\boldsymbol{y} = \boldsymbol{B}\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{y} \end{bmatrix} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{f} \in \mathbb{R}^{d_f}$ denotes the latent confounder, which can affect the multiple observed variables $\boldsymbol{y}$, and $\boldsymbol{B} \in \mathbb{R}^{d_y \times (d_f + d_y)}$ denotes the causal matrix collecting the connection strengths b_{ij} from f_i or y_i to y_j.

2.2 Counterfactual Inference

After estimating the structural equations, you can infer counterfactual quantities. Pearl [32] described counterfactual inference to any causal model as three steps.

Definition 2 (Counterfactual Inference)**.** *Given the evidence w, to compute the probability of $X = x$ under the hypothetical condition $S = s$ (S is a subset of variables), counterfactual inference proceeds in the following three steps:*

Step 1 (abduction): *Update the prior distribution $P(u)$ to the posterior distribution $P(u|w)$.*

Step 2 (action): *Replace the equations for S with the equations $S = s$.*

Step 3 (prediction): *Compute the distribution on the remaining equations and obtain the probability of $X = x$.*

Step 2 in Definition 2 is also called an intervention in S. In these three steps, the amount of change in the probability of $X = x$ is called a causal effect of S on X, which can be interpreted as the amount indicating how much S affects X.

2.3 Counterfactual Fairness

Kusner *et al.* [25] focus on the causal effect and define counterfactual fairness. We assume that a classifier and attributes are binary without loss of generality and give the definition below.

Definition 3 (Counterfactual Fairness)**.** *Let $\hat{Y}$ denote a prediction of a binary classifier, S denote a binary sensitive attribute, and $Z \subseteq X$ denote a set of attributes (X is a set of non-sensitive attributes). The classifier satisfies counterfactual fairness if we have*

$$P(\hat{Y}_{S \leftarrow s} | S = s, Z = z) = P(\hat{Y}_{S \leftarrow s'} | S = s, Z = z) \tag{3}$$

under any condition $Z = z$, where $s, s' \in \{0, 1\}$, and $\hat{Y}_{S \leftarrow s}$ denotes the variable $\hat{Y}$ after the intervention that replaces the equations for S with $S = s$.

Here, the sensitive attribute is some traits identified by law [31] on which it is illegal to discriminate against.

Definition 3 is based on the belief that the classifier should make the same prediction both in the actual world and the counterfactual world. In other words, the causal effect of the sensitive attribute on the output of the classifier should be zero.

3. CUV-GAN

In this section, we introduce CUV-GAN, which provides a method of generating counterfactual images. CUV-GAN is a network architecture based on causal relationships including attributes and images.

3.1 Causal Model of CUV-GAN

We consider the causal relationships including images as the causal diagram in Fig. 2 (a). The generation process of the attribute variables $\boldsymbol{y}$ followed LvLiNGAM [18]. In the generation process of the image $\boldsymbol{x}$, we consider any factors that affect the image $\boldsymbol{x}$ as the latent confounders $\boldsymbol{f}$ and the attribute variables $\boldsymbol{y}$. Hence, we derive the structural equation for the image $\boldsymbol{x}$ as follows:

$$\boldsymbol{x} = f_x(\boldsymbol{f}, \boldsymbol{y}), \tag{4}$$

where $f_x : \mathcal{F} \times \mathcal{Y} \to \mathcal{X}$. Here, $\mathcal{F} \subseteq \mathbb{R}^{d_f}$ denotes the latent confounders space, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denotes the attribute variables space, and $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ denotes the image space. $\boldsymbol{f}$ and $\boldsymbol{y}$ are according to U_{pa_i} and pa_i in Definition 1.

Since the structural equations (2) (4) are hardly applicable to image generation directly, we employed a GAN's generator [14], which is a powerful technique for image generation. The well-trained generator can be interpreted as the deterministic function $G : \mathcal{Z} \to \mathcal{X}$ where $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ denotes the latent space of GAN. The latent vector has all information of the image. Hence, we make the following assumptions about the latent vector and the variables that determine the image.

Assumption 1. *There is a one-to-one correspondence between the latent vector of GAN and the variables that determine the image.*

As long as Assumption 1 holds, the relation between them can be described as an invertible function. Since we consider that the variables that determine the image are the latent confounders $\boldsymbol{f}$ and the attribute variables $\boldsymbol{y}$, the relation can be written as an invertible function $H : \mathcal{F} \times \mathcal{Y} \to \mathcal{Z}$. Therefore, we can formulate the structural equations for the image $\boldsymbol{x}$ as follows:

$$\boldsymbol{z} = H(\boldsymbol{f}, \boldsymbol{y}), \tag{5}$$

$$\boldsymbol{x} = G(\boldsymbol{z}). \tag{6}$$

Given the above (2) (5) (6), the causal diagram of CUV-GAN is shown in Fig. 2 (b).

3.2 Training Procedure of CUV-GAN

We assume that a well-trained generator is given, otherwise, you need to train the generator according to any GAN frameworks [2] [21] [22].

CUV-GAN aims to generate counterfactual images of real images for improving counterfactual fairness. To embed the real images into the latent space of GAN, you can use GAN-Inversion [5] [16] [29] [39] [40], which is a technique for inferring the latent vector from the image.

LvLiNGAM (2) can be transformed as follows:

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{y} \end{bmatrix} + \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{e} \end{bmatrix}, \tag{7}$$

$$\left(\boldsymbol{I} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{e} \end{bmatrix}, \tag{8}$$

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{y} \end{bmatrix} = \left(\boldsymbol{I} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix}\right)^{-1} \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{e} \end{bmatrix}, \tag{9}$$

$$= \tilde{\boldsymbol{A}} \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{e} \end{bmatrix}, \tag{10}$$

$$\left(\tilde{\boldsymbol{A}} = \left(\boldsymbol{I} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix}\right)^{-1}\right),$$

$$\boldsymbol{y} = \boldsymbol{A} \begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{e} \end{bmatrix}, \tag{11}$$

where $\boldsymbol{A}$ is the bottom d_y rows of $\tilde{\boldsymbol{A}}$. Hence, you can solve LvLiNGAM as overcomplete independent component analysis (OICA). Note that you can calculate the causal matrix $\boldsymbol{B}$ from the matrix $\boldsymbol{A}$ by using the following equation.

$$\tilde{\boldsymbol{A}} = \left(\boldsymbol{I} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{B} \end{bmatrix}\right)^{-1}, \tag{12}$$

where $\tilde{\boldsymbol{A}}$ is the concatenation of the following matrix $(a_{ij})_{ij} (\in \mathbb{R}^{d_f \times d_f + d_y})$ on the top of $\boldsymbol{A}$.

$$a_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (otherwise) \end{cases} \tag{13}$$

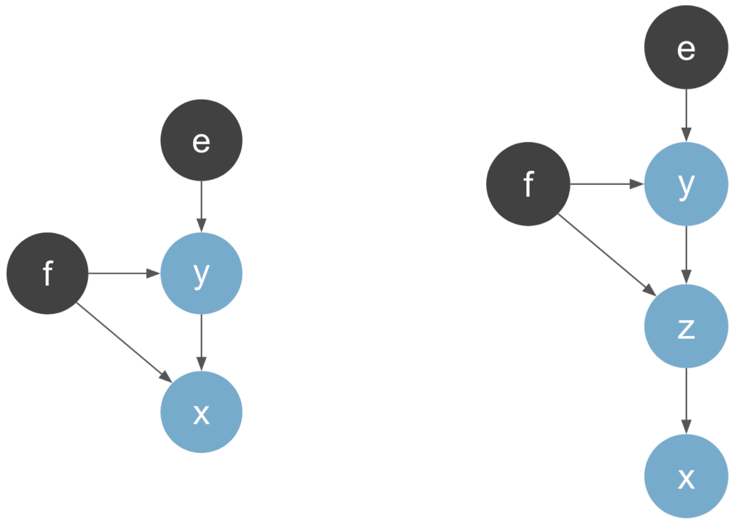
To train the invertible function H, you can use Flow [8] [9] [15]. Let the well-trained generator, the pre-trained encoder, and the inferred matrix be G, E, and $\boldsymbol{A}$, then we learn the parameter of the invertible function θ_H by the following optimization problem:

$$\theta_H^* = \underset{\theta_H}{\arg\max} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y}) \sim \mathcal{D},\ \hat{\boldsymbol{e}} \sim P(\hat{\boldsymbol{e}})} \left[L\left(\hat{\boldsymbol{y}}, \boldsymbol{y}\right) + \lambda L\left(\boldsymbol{A} \begin{bmatrix} \hat{\boldsymbol{f}} \\ \hat{\boldsymbol{e}} \end{bmatrix}, \boldsymbol{y}\right)\right], \tag{14}$$

$$\text{where} \begin{bmatrix} \hat{\boldsymbol{f}} \\ \hat{\boldsymbol{y}} \end{bmatrix} = H^{-1}\left(E(\boldsymbol{x})\right), \tag{15}$$

$$\lambda \in \mathbb{R}_+. \tag{16}$$

Here, $P(\hat{\boldsymbol{e}})$ is the distribution of the exogenous variables learned by OICA. L is the loss function for evaluating the difficulty between the true value and the predicted value such as a mean squared error. λ is the hyperparameter for a positive real number such as $d_f/(d_f + d_y)$.



(a) Causal diagram with images (b) Causal diagram with images using latent vectors

Fig. 2 Causal Diagrams with Images. $\boldsymbol{e}$, $\boldsymbol{f}$, $\boldsymbol{y}$, $\boldsymbol{z}$, $\boldsymbol{x}$ denote exogenous variables, latent confounders, attribute variables, latent variables of the well-trained generator, and images, respectively. Black nodes and blue nodes denote unobserved variables and observed variables.

3.3 Generating Counterfactual Images

After the training, we can generate the counterfactual image from the real image. Suppose our purpose is the intervention to change $\boldsymbol{s} \in \boldsymbol{y}$ to $\boldsymbol{s'}$. First, we embed the real image by using the pre-trained encoder E and obtain the inferred latent vector $\hat{\boldsymbol{z}}$. Next, we input the latent vector $\hat{\boldsymbol{z}}$ into the inverse function of the pre-trained invertible function H^{-1} and obtain the inferred attribute variables $\hat{\boldsymbol{y}}$ and the inferred latent confounders $\hat{\boldsymbol{f}}$. Then, we infer the exogenous variables $\hat{\boldsymbol{e}}$ from Equation (8) using the inferred causal matrix $\boldsymbol{B}$. Then, we replace the structural equations for $\boldsymbol{s}$ with $\boldsymbol{s} = \boldsymbol{s'}$, calculate the structural equations for the remaining attribute variables $\boldsymbol{y}\backslash\boldsymbol{s}$, and obtain a counterfactual attribute variables $\boldsymbol{y_{s\leftarrow s'}}$. Finally, we generate a counterfactual image $G(H(\hat{\boldsymbol{f}}, \boldsymbol{y_{s\leftarrow s'}}))$.

These procedures are shown in Algorithm 1. We can confirm that our Algorithm 1 is according to Definition 2.

4. Experiments

4.1 Dataset and Implementation

To evaluate the generating quality of CUV-GAN, we created the synthetic dataset that has 10,000 simple facial images with the attribute variables. The attributes and images in our dataset were generated according to our assumption about the generation process. First, the exogenous variables and the latent confounders were sampled from their distributions. Second, the observed variables were generated from the parent variables. Finally, the images were drawn from the observed variables and the latent confounders. We show examples of our dataset in Figure 3 and the causal diagram of the observed variables and the latent confounders in Figure 4. In detail, the structural equations for the attributes are as follows:

Algorithm 1 Generating Counterfactual Images with Interventions in $\boldsymbol{s}$

```
Input: Image $\boldsymbol{x}$, attributes $\boldsymbol{y}$, intervention variables $\boldsymbol{s} \subseteq \boldsymbol{y}$,
    trained generator $G$, trained encoder $E$,
    inferred causal matrix $\boldsymbol{B}$, trained invertible function $H$.
Output: counterfactual image $\hat{\boldsymbol{x}}$

    Abduction :
$\hat{\boldsymbol{z}} = E(\boldsymbol{x})$.
$\begin{bmatrix}\hat{\boldsymbol{f}}\\ \hat{\boldsymbol{y}}\end{bmatrix} = H(\hat{\boldsymbol{z}})$.
$\begin{bmatrix}\hat{\boldsymbol{f}}\\ \hat{\boldsymbol{e}}\end{bmatrix} = \left(\boldsymbol{I} - \begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{B}\end{bmatrix}\right)\begin{bmatrix}\hat{\boldsymbol{f}}\\ \boldsymbol{y}\end{bmatrix}$.

    Action :
Set $\boldsymbol{s} = \boldsymbol{s'}$

    Prediction :
Let the indices of $\boldsymbol{s}$ be $\boldsymbol{i_s}$.
for $i \leftarrow \{0, ..., d_y\}\backslash\boldsymbol{i_s}$ do
    $y_i \leftarrow \boldsymbol{b}_i \begin{bmatrix}\hat{\boldsymbol{f}}\\ \boldsymbol{y}\end{bmatrix} + e_i, (\boldsymbol{b}_i$: the i-th row of $\boldsymbol{B}$).
end for
Let the attribute variables after the intervention be $\boldsymbol{y_{s\leftarrow s'}}$.
$\boldsymbol{z_{s\leftarrow s'}} = H^{-1}\left([\hat{\boldsymbol{f}}, \boldsymbol{y_{s\leftarrow s'}}]\right)$.
$\boldsymbol{x_{s\leftarrow s'}} = G(\boldsymbol{z_{s\leftarrow s'}})$.
```

$$f_r = e_r,\ e_r \sim \mathcal{U}(0,1), \tag{17}$$
$$y_s = e_s,\ e_s \sim \mathcal{U}(0,1), \tag{18}$$
$$y_c = 0.4f_r + e_c,\ e_c \sim \mathcal{U}(0,0.6), \tag{19}$$
$$y_e = 0.2f_r + 0.3y_s + e_e,\ e_e \sim \mathcal{U}(0,0.5), \tag{20}$$

where $\mathcal{U}$ is the continuous uniform distribution.

Since the structural equation for the image is extremely complex, we can not write down the equation. Intuitively, the eyes attribute controls the diameter of the eyes in the image, the smiling attribute controls the curve of the mouth,

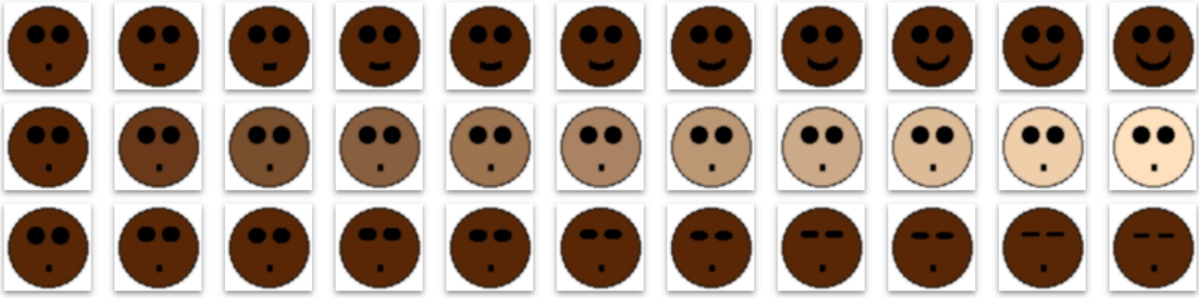


Fig. 3 Examples of Our Dataset. The images are determined by attributes. The smiling attribute determines the degree of the mouth curvature. The skin attribute determines the color of the skin. The eyes attribute determines the size of the eyes. From the top row, this figure illustrates the results of changing the smiling, color, eyes attributes from 0 to 1.

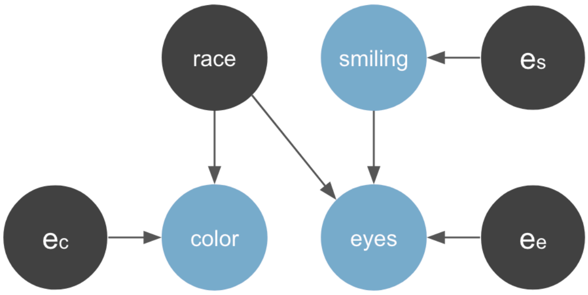


Fig. 4 Causal Diagram of Our Dataset. The smiling, color, and eyes attributes are the observed variables. The race attribute is the latent confounder. We modeled simple relationships in the human face as follows: Race affects skin color and eyes size. Smiling affects the degree of eyes openness.

and the color attribute controls RGB values inside the circle of the image.

Although we assume that a well-trained generator is given, we need to train the generator for the self-making dataset. In this paper, we employed "Mode-Seeking Generative Adversarial Network" (MSGAN) [30] as the framework for training the generator. MSGAN explicitly maximizes the ratio of the distance between generated images with respect to the corresponding latent vectors, thus preventing the generator from the mode collapse issue.

Since the pre-trained encoders such as Multi-Code GAN Inversion [16] and Lia [40] are available online, you do not need to train the encoder in a wide range of situations. However, we need to train the encoder for the self-making dataset. In this paper, we employed "In-Domain GAN Inversion" (Idinvert) [39] as the framework for training the encoder. Idinvert embeds the images into the latent vectors that minimize the distance between the original images and the reconstructed images in the image space and the feature space.

To solve LvLiNGAM (2), we employed "Likelihood-Free Overcomplete Independent Component Analysis" (LFOICA) [7]. LFOICA explores the matrix $\boldsymbol{A}$ in OICA (11) by stochastic gradient descent and learns neural networks that transform Gaussian distributions to the distributions of the independent components. Then, we use Gaussian distribution and the trained neural networks as the prior distribution of the exogenous variables $\boldsymbol{e}$ for training the invertible function.

To train the invertible function, we employed "Non-linear Independent Components Estimation" (NICE) [8]. NICE estimates the invertible transformation of the distributions using additive coupling layers.

4.2 Qualitative Evaluation

We trained CUV-GAN with our dataset and generated the counterfactual images by Algorithm 1. The results are shown in Figure 5. We show the generated images by CUV-GAN and also "Matrix Subspace Projection" (MSP) [27]. The images for each row in Figure 5 are generated by performing the continuous interventions in the attribute from 0 to 1 in one sample.

CUV-GAN generated images of eyes more smoothly changing and represented darker colors when the value of the intervention is close to 0 than MSP. While MSP changed only the smiling attribute, CUV-GAN changed the smiling attribute and also the eyes attribute, which is a child node of the smiling attribute in the causal diagram. While MSP could not change the latent confounders, CUV-GAN could perform the intervention in it. Although CUV-GAN did not explicitly perform the intervention in the race attribute, it can be seen that the color and eyes attributes were changed as a result.

4.3 Quantitative Evaluation

We built a regressor that predicts the attribute values from the input image to quantitatively evaluate CUV-GAN. As a result of training, the regressor has been able to predict for each attribute with a mean squared error of 0.0015 in the test dataset. We quantitatively evaluated the generated counterfactual images by calculating mean squared errors between the ground truth of the counterfactual attributes and the outputs when inputting the generated counterfactual images into the regressor. The results are shown in Table 1. CUV-GAN generated higher quality images than MSP [27].

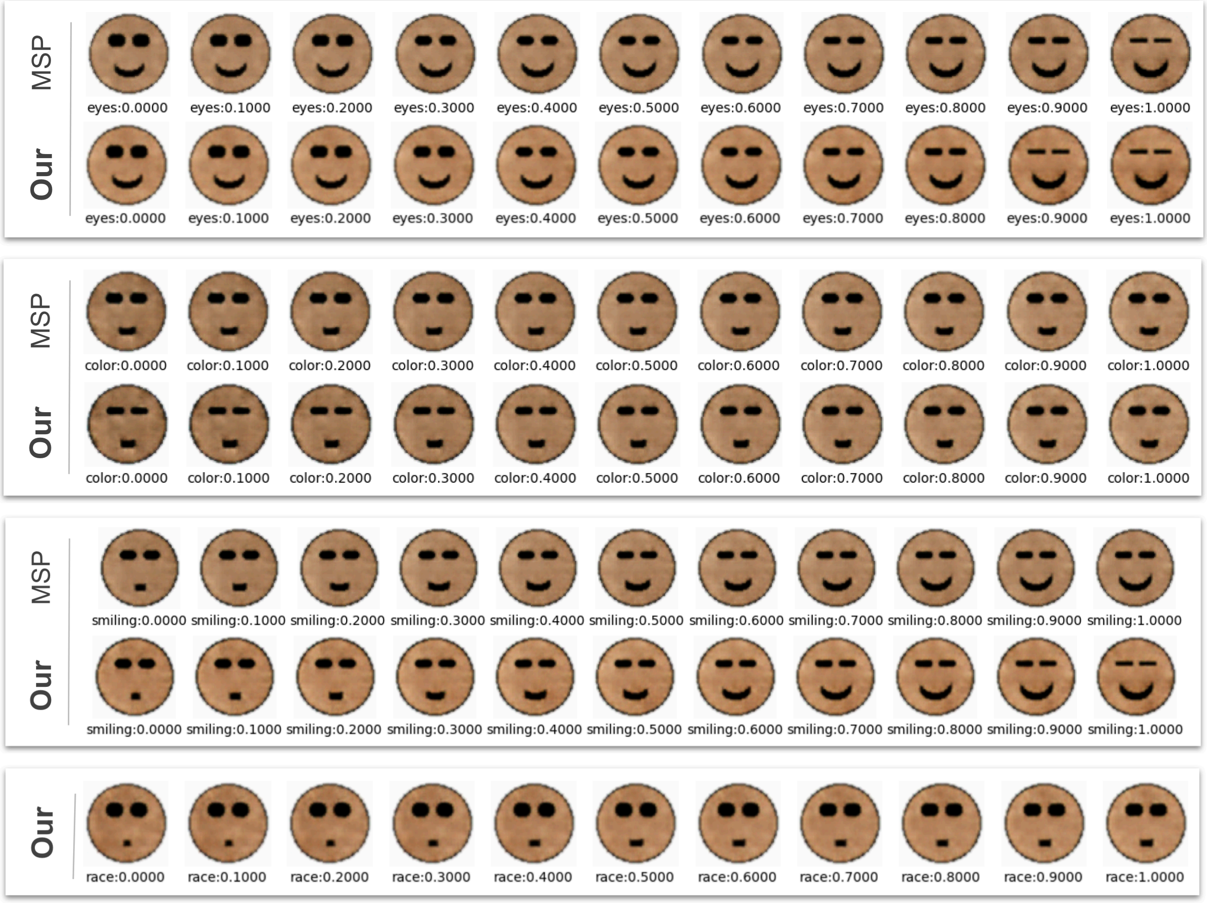

MSP
eyes:0.0000 eyes:0.1000 eyes:0.2000 eyes:0.3000 eyes:0.4000 eyes:0.5000 eyes:0.6000 eyes:0.7000 eyes:0.8000 eyes:0.9000 eyes:1.0000
Our
eyes:0.0000 eyes:0.1000 eyes:0.2000 eyes:0.3000 eyes:0.4000 eyes:0.5000 eyes:0.6000 eyes:0.7000 eyes:0.8000 eyes:0.9000 eyes:1.0000
MSP
color:0.0000 color:0.1000 color:0.2000 color:0.3000 color:0.4000 color:0.5000 color:0.6000 color:0.7000 color:0.8000 color:0.9000 color:1.0000
Our
color:0.0000 color:0.1000 color:0.2000 color:0.3000 color:0.4000 color:0.5000 color:0.6000 color:0.7000 color:0.8000 color:0.9000 color:1.0000
MSP
smiling:0.0000 smiling:0.1000 smiling:0.2000 smiling:0.3000 smiling:0.4000 smiling:0.5000 smiling:0.6000 smiling:0.7000 smiling:0.8000 smiling:0.9000 smiling:1.0000
Our
smiling:0.0000 smiling:0.1000 smiling:0.2000 smiling:0.3000 smiling:0.4000 smiling:0.5000 smiling:0.6000 smiling:0.7000 smiling:0.8000 smiling:0.9000 smiling:1.0000
Our
race:0.0000 race:0.1000 race:0.2000 race:0.3000 race:0.4000 race:0.5000 race:0.6000 race:0.7000 race:0.8000 race:0.9000 race:1.0000

Fig. 5 Generated Counterfactual Images. For each row, the images are the results of the interventions that continuously change an attribute value from 0 to 1 in one sample. From the top row, the images are the results of the interventions in the eyes, color, smiling, and race attributes. Since the race attribute is the unobserved variables, CUV-GAN can not explicitly perform the intervention in the race. However, as a result, we can see that the interventions in the latent confounder change the color and eyes attributes.

	eyes	color	smiling	race
MSP [27]	0.0698	0.0671	0.0728	-
CUV-GAN	**0.0046**	**0.0050**	**0.0049**	**0.0364**

Table 1 Quantitative evaluation of the generated counterfactual images. The numbers in the table denote mean squared errors between the ground truth of counterfactual attributes and the outputs when inputting the generated images into the trained regressor that can predict attributes from the images.

DA	Acc. ↑	DP ↓	EO ↓
	0.9647 ± 0.0032	0.0176 ± 0.0019	0.0138 ± 0.0035
✓	**0.9679 ± 0.0029**	**0.0145 ± 0.0042**	**0.0108 ± 0.0037**

Table 2 Comparison of accuracy and fairness in cases using and not using the generated counterfactual images of CUV-GAN as data augmentation. DA, Acc., DP, and EP denotes data augmentation, accuracy, demographic parity, and equal opportunity. The numbers in bold indicate better performance than another.

4.4 Assessing Fairness

From Definition 3, we considered that a model becomes fair by leaning the counterfactual images in the sensitive attribute. We trained the simple convolutional neural network (CNN) which predicts whether the input image is smiling as a baseline model. Next, We used the resultant images of the intervention in the color attribute as data augmentation for retraining the baseline model. The results are shown in Table 2. CNN became fair by using the generated counterfactual images.

5. Conclusion

We assumed a generation process of the images as the causality and formulated structural equations using the well-trained generator of GAN. We proposed CUV-GAN, which is a framework based on the causal equations and can perform interventions and generate counterfactual images. In qualitative and quantitative experiments using an original synthetic dataset, it was confirmed that CUV-GAN can generate the counterfactual images and be useful for improving fairness in machine learning.

As future works, we consider experiments with real datasets such as CelebA dataset [28] and assessing counterfactual fairness of various datasets and image classifiers.

Acknowledgments This study was partially supported by the MIC/SCOPE #172107101, JSPS KAKENHI (19H04172, 19K20344), and JST-Mirai Program (JPMJMI18B4), Japan.

References

[1] Bollen, K.: *Structural Equations with Latent Variables*, Wiley-Interscience (1989).

[2] Brock, A., Donahue, J. and Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis, *arXiv preprint arXiv:1809.11096* (2018).

[3] Chen, J., Kallus, N., Mao, X., Svacha, G. and Udell, M.: Fairness Under Unawareness, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, (online), DOI: 10.1145/3287560.3287594 (2019).

[4] Chiappa, S.: Path-Specific Counterfactual Fairness, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 01, pp. 7801–7808 (online), DOI: 10.1609/aaai.v33i01.33017801 (2019).

[5] Creswell, A. and Bharath, A. A.: Inverting the generator of a generative adversarial network, *IEEE transactions on neural networks and learning systems*, Vol. 30, No. 7, pp. 1967–1974 (2018).

[6] Denton, E., Hutchinson, B., Mitchell, M., Gebru, T. and Zaldivar, A.: Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias (2020).

[7] Ding, C., Gong, M., Zhang, K. and Tao, D.: Likelihood-Free Overcomplete ICA and Applications in Causal Discovery (2019).

[8] Dinh, L., Krueger, D. and Bengio, Y.: Nice: Non-linear independent components estimation, *arXiv preprint arXiv:1410.8516* (2014).

[9] Dinh, L., Sohl-Dickstein, J. and Bengio, S.: Density estimation using Real NVP (2017).

[10] Dua, D. and Graff, C.: UCI Machine Learning Repository (2017).

[11] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R.: Fairness Through Awareness (2011).

[12] Fernandes, J. A., Irigoien, X., Lozano, J. A., Inza, I., Goikoetxea, N. and Pérez, A.: Evaluating machine-learning techniques for recruitment forecasting of seven North East Atlantic fish species, *Ecological Informatics*, Vol. 25, pp. 35–42 (2015).

[13] Ferryman, K. and Pitcan, M.: Fairness in precision medicine (2018).

[14] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative Adversarial Networks (2014).

[15] Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I. and Duvenaud, D.: Ffjord: Free-form continuous dynamics for scalable reversible generative models, *arXiv preprint arXiv:1810.01367* (2018).

[16] Gu, J., Shen, Y. and Zhou, B.: Image Processing Using Multi-Code GAN Prior (2020).

[17] Hardt, M., Price, E. and Srebro, N.: Equality of Opportunity in Supervised Learning (2016).

[18] Hoyer, P. O., Shimizu, S., Kerminen, A. J. and Palviainen, M.: Estimation of causal effects using linear non-Gaussian causal models with hiddden variables, *International Journal of Approximate Reasoning* (2008).

[19] Jahanian, A., Chai, L. and Isola, P.: On the "steerability" of generative adversarial networks (2020).

[20] Javier, S. M., Lina, D. and Lilian, E.: What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems, *CoRR*, Vol. abs/1910.06144 (online), available from ⟨http://arxiv.org/abs/1910.06144⟩ (2019).

[21] Karras, T., Aila, T., Laine, S. and Lehtinen, J.: Progressive Growing of GANs for Improved Quality, Stability, and Variation (2018).

[22] Karras, T., Laine, S. and Aila, T.: A style-based generator architecture for generative adversarial networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410 (2019).

[23] Kenneth, H., Vaughan, J. W., III, H. D., Miroslav, D. and Wallach, H. M.: Improving fairness in machine learning systems: What do industry practitioners need?, *CoRR*, Vol. abs/1812.05239 (online), available from ⟨http://arxiv.org/abs/1812.05239⟩ (2018).

[24] Komatsu, Y., Shimizu, S. and Shimodaira, H.: Assessing statistical reliability of LiNGAM via multiscale bootstrap, *International Conference on Artificial Neural Networks*, Springer, pp. 309–314 (2010).

[25] Kusner, M. J., Loftus, J., Russell, C. and Silva, R.: Counterfactual Fairness, *Advances in Neural Information Processing Systems* (Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. and Garnett, R., eds.), Vol. 30, Curran Associates, Inc., pp. 4066–4076 (2017).

[26] Lai, P.-C. and Bessler, D. A.: Price discovery between carbonated soft drink manufacturers and retailers: a disaggregate analysis with PC and LiNGAM algorithms, *Journal of Applied Economics*, Vol. 18, No. 1, pp. 173–197 (2015).

[27] Li, X., Lin, C., Li, R., Wang, C. and Guerin, F.: Latent Space Factorisation and Manipulation via Matrix Subspace Projection (2020).

[28] Liu, Z., Luo, P., Wang, X. and Tang, X.: Deep Learning Face Attributes in the Wild, *Proceedings of International Conference on Computer Vision (ICCV)* (2015).

[29] Ma, F., Ayaz, U. and Karaman, S.: Invertibility of Convolutional Generative Networks from Partial Measurements, *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. and Garnett, R., eds.), Vol. 31, Curran Associates, Inc., pp. 9628–9637 (2018).

[30] Mao, Q., Lee, H.-Y., Tseng, H.-Y., Ma, S. and Yang, M.-H.: Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis (2019).

[31] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A.: A Survey on Bias and Fairness in Machine Learning (2019).

[32] Pearl, J.: *Causality: Models, Reasoning and Inference*, Cambridge University Press (2009).

[33] Perarnau, G., van de Weijer, J., Raducanu, B. and Álvarez, J. M.: Invertible Conditional GANs for image editing (2016).

[34] Reisig, M. D., Holtfreter, K. and Morash, M.: Assessing Recidivism Risk Across Female Pathways to Crime, *Justice Quarterly*, Vol. 23, No. 3, pp. 384–405 (online), DOI: 10.1080/07418820600869152 (2006).

[35] Shen, Y., Yang, C., Tang, X. and Zhou, B.: InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs (2020).

[36] Shimizu, S., Hoyer, P. O., Hyvarinen, A. and Kerminen, A.: A Linear Non-Gaussian Acyclic Model for Causal Discovery, *Journal of Machine Learning Research 2006* (2006).

[37] Wright, S.: Colrrelation and causation, *Journal of Agricultural Research* (1921).

[38] Wu, Y., Zhang, L. and Wu, X.: Counterfactual Fairness: Unidentification, Bound and Algorithm, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, International Joint Conferences on Artificial Intelligence Organization, pp. 1438–1444 (online), DOI: 10.24963/ijcai.2019/199 (2019).

[39] Zhu, J., Shen, Y., Zhao, D. and Zhou, B.: In-Domain GAN Inversion for Real Image Editing (2020).

[40] Zhu, J., Zhao, D., Zhang, B. and Zhou, B.: Disentangled Inference for GANs with Latently Invertible Autoencoder (2020).