Article

Fair Facial Attribute Classification via Causal Graph-Based Attribute Translation

Sunghun Kang *, Gwangsu Kim and Chang D. Yoo *

School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea; sunghun.kang@kaist.ac.kr (S.K.); s88012@kaist.ac.kr (G.K.)
* Correspondence: cd_yoo@kaist.ac.kr

Abstract: Recent studies have raised concerns regarding racial and gender disparity in facial attribute classification performance. As these attributes are directly and indirectly correlated with the sensitive attribute in a complex manner, simple disparate treatment is ineffective in reducing performance disparity. This paper focuses on achieving counterfactual fairness for facial attribute classification. Each labeled input image is used to generate two synthetic replicas: one under factual assumptions about the sensitive attribute and one under counterfactual. The proposed causal graph-based attribute translation generates realistic counterfactual images that consider the complicated causal relationship among the attributes with an encoder–decoder framework. A causal graph represents complex relationships among the attributes and is used to sample factual and counterfactual facial attributes of the given face image. The encoder–decoder architecture translates the given facial image to have sampled factual or counterfactual attributes while preserving its identity. The attribute classifier is trained for fair prediction with counterfactual regularization between factual and corresponding counterfactual translated images. Extensive experimental results on the CelebA dataset demonstrate the effectiveness and interpretability of the proposed learning method for classifying multiple face attributes.

Keywords: deep learning; fair classification; facial attribute classification

1. Introduction

Facial attribute classification detects the presence versus absence of various labeled attributes, including bio-metric features (“big nose”), expression (“smiling”), and worn accessories (“glasses”). Attribute classification supports various tasks including tagging, searching, detecting, and verification of identity [1–8]. Existing facial attribute classification algorithms have solely been focused on prediction accuracy and are highly likely to suffer from prediction bias that leads to disparity in performance among various population subgroups belonging to different gender, race and age. In fact, recent studies have exposed the racial and gender disparity in the performance of a number of commercial face recognition systems. For public and commercial use, prediction must be fair.

Although considerable research in AI has been devoted to mitigating general prediction bias for fair prediction, a substantial portion of the research had been centered on attaining group fairness [9–13] which is focused on reducing performance disparity among groups with distinct sensitive attributes. This notion of group fairness does not necessarily equate well with the interest of an individual [12,14]. For any system that affects the welfare or opportunities available to an individual, individual fairness should be considered over group fairness. The counterfactual fairness is an attractive alternative as it captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group. It requires the probability distribution for the individual’s classifier output label to be unchanged regardless of the switched value of the sensitive attribute [15].
With the aid of deep learning [16–19], face recognition performance has substantially improved to the point of exhibiting close to human performance. However, a number of studies [20,21] have noted disparity in prediction performance among groups with different sensitive attributes that include gender and race. It is observed that prediction accuracy is lower for African-Americans and females than for Caucasians and males. In an attempt to resolve this issue, a method based on unsupervised domain adaptation IMAN [20] has been proposed to improve recognition performance for ethnic minority groups.

One common cause for prediction bias comes as a consequence of a trained model inheriting the bias in the training dataset [22]: the sampling bias and societal bias [23,24] in the dataset lead to skewed distributions with respect to the sensitive attribute. One measure to counteract this bias inheritance is to collect a more balanced dataset. FairFace [25] with an emphasis on balanced mixed-race composition was released for developing classification models whose accuracy is consistent between race and gender groups. However, collecting a dataset with balanced facial attributes is a challenging task for a number of reasons including (1) societal bias, e.g., younger people, more than seniors, take photos and upload them on the web, and (2) the complex relationship that exists among facial attributes, leading to a greater-than-exponential rise in the number of training images that must be collected in order to obtain constant minimum sample sizes for all combinations of a linearly increasing number of attributes. These factors make fair data collection nearly impossible.

To achieve fair face-attribute classification, an algorithm is required that mitigates bias regardless of the sample imbalance and the existence of complex relationships among attributes. This paper studies a learning method that focuses on achieving counterfactual fairness for face-attribute classification based on an encoder–decoder framework that translates the input image into both factual and counterfactual images using a causal graph of the attributes. More specifically, from the facial attribute labels in the training dataset, the complex relationships between the facial attributes and the sensitive attribute are modeled and represented by a causal graph. Observed input facial attributes seed a sampling algorithm that generates the attribute lists for faces with factual and counterfactual settings of the sensitive attribute, based on the discovered causal graph. The causal graph-based attribute translator generates realistic factual and counterfactual images corresponding to the observed image and sampled counterfactual facial attributes. The attributes of the generated counterfactual image follow the intervened attribute distribution represented by the causal graph. Based on the factual and counterfactual image and label pair, a counterfactual regularization term is proposed that penalizes differences between the prediction probabilities of the counterfactual and factual images, thereby encouraging the learning of a counterfactually fair face attribute classifier. The counterfactual regularization penalizes prediction differences only for attributes with the same sampled value, thereby learning fairness without the need to build a separate causal graph for each attribute.

Our contributions can be summarized as follows:

- An encoder–decoder framework is proposed, incorporating a causal graph of face attributes, to facilitate input image translation in generating factual and counterfactual images for counterfactually fair learning.
- A counterfactual regularization term is incorporated to penalize counterfactually unfair facial attribute classification, thereby reducing the counterfactual fairness disparity in multi-label classification.
- We demonstrate the overall framework can achieve counterfactual fairness on the CelebA corpus, and we also provide factual and counterfactual images with multi-dimensional attributes induced from the causal graph.
- This work, to the best of our knowledge, is the first attempt in providing counterfactual fairness considering the complicated causal relationships among face attributes, including sensitive attributes, for fair facial-attribute classification.

The remainder of this paper is organized as follows. Section 2 briefly reviews some of the most relevant literature related to the proposed framework. Section 3 describes counterfactual fairness for face attribute classification. Section 4 describes the proposed learning
method to achieve counterfactual fairness for face attribute classification. Experimental results are reported in Section 5. Section 6 provides a summary and conclusion of the paper.

2. Related Work

2.1. Fairness in Computer Vision

A number of studies [21,26] have raised concerns regarding unfair prediction in commercialized recognition systems, and a number of studies [20,27–32] have been conducted to achieve fair prediction on various computer vision tasks. Several image captioning and VQA models [27,33,34] have been noted to exaggerate biases [27,34] that lead to incorrect captions due to over-reliance on prejudicial context. To mitigate bias, the models focus on the object, not the context. In the image classification task, the Blind network [29] introduces confusion loss, which penalizes non-uniformly distributed predictions of the sensitive attribute in order to learn fair representation. Domain adaptation is one of the methods to reduce the performance gap between different ethnic groups [20]. The information of the major groups can be transferred to minor groups for reducing the performance gap.

2.2. Counterfactual Fairness

In contrast to other group fairness measures, the counterfactual fairness measure [35–37] is grounded on the causal relationship among variables. The counterfactual fairness criterion states that a prediction is fair toward an individual if the predictions of the actual world and counterfactual world are the same regardless of the sensitive attribute. The counterfactual is obtained by intervention in the causal graph. The origin of biased prediction can be analyzed using the causal graph represented as a Bayesian network [38]. It provides a graphical interpretation of unfairness in a dataset as the presence of an unfair causal path in the causal Bayesian network. Chiappa and Issac [39] analyze the fair decision system in complex scenarios where a sensitive attribute might affect the decision along both fair and unfair pathways in a causal model.

For computer vision tasks, Gender Slopes [40] are proposed for evaluating counterfactual fairness. Gender Slopes synthesize counterfactual images using an encoder–decoder framework, then measure the counterfactual fairness gap. The study of [40] reveals that state-of-the-art commercial computer vision classifiers produce biased predictions against gender attributes. However, the Gender Slopes do not consider the causal relationship between variables including the sensitive attribute. In facial attribute classification, Denton et al. [41] proposed measuring counterfactual fairness by generating a facial image with a Generative Adversarial Network (GAN). Here, the effect of an orthogonal vector to the decision boundary of each attribute classifier on a one-dimensional smiling classifier is investigated.

2.3. Image Generation

Goodfellow et al. [42] proposed a framework for training a deep generative model in an adversarial manner referred to as the generative adversarial network (GAN). This framework simultaneously trains two different networks to mimic the data distribution: (1) a discriminator is designed to distinguish between the generated and original sample, and (2) a generator is trained to generate samples realistic enough to fool the discriminator. Building on the success of the GAN, various studies have been conducted to generate more realistic images. PGGAN [43] is proposed to increase the image resolution progressively during the training. Recently, BigGAN [44] is proposed to generate more realistic images with deeper architecture. Causal GAN [45] is proposed to sample both the observational and intervention distribution. They use an additional GAN for controlling the causal relationship.

There are also some conditional GANs [46,47] to generate new images from given images. AttGAN [48] is one of the most successful architectures for generating facial images with discretely modified facial attributes. The additional classifier, which predicts the facial
attributes of generated images, provides supervision indicating which attributes should appear in the generated facial images.

This paper differs from CausalGAN and AttGAN in that: the proposed causal graph-based attribute translator generates the counterfactual images from a given image and facial attributes. The CausalGAN [45] is proposed to generate facial images considering the causal relationship among the attributes; the CausalGAN can sample facial images from the intervened distribution of the causal relationship discovered from the training facial image and attribute vector pairs. However, CausalGAN does not guarantee the generated facial images keep the identity of the given images, and it cannot generate counterfactual images. To generate counterfactual face images of a given observed sample, a causal graph that explains the data distribution is required. The conditional GAN algorithms [44,46–48], including AttGAN, require the desired attribute to generate the given facial image without consideration of the causal relationship. Without a causal graph that describes the relationship between the attributes and the image, the conditional generation algorithm cannot generate counterfactual facial images.

3. Problem Definition

Given facial image \( x \in X \subset \mathbb{R}^{c \times w \times h} \), where \( c, w, \) and \( h \) are the dimensions of channel, width, and height, respectively, our task is to predict binary ground truth label \( y = (y_1, y_2, \ldots, y_M) \) where \( y_i \in \{0,1\} \) for \( i = 1, \ldots, M \) representing the presence of each of the \( M \) attributes. Here, the elements of \( y \) are correlated. A classifier parameterized by \( \theta, f_\theta \), is learned to predict the presence of \( M \) facial attributes as \( \hat{y} = I(\hat{p} > 0.5) \), where \( \hat{p} = f_\theta(x) \) and \( I \) is an indicator function. Here, \( f_\theta \) is assumed to be a convolutional neural network (CNN), and \( \theta \) is learned to minimize the summation of cross-entropy losses between the ground-truth attribute vector \( y_i \) and \( \hat{y}_i, i = 1, \ldots, M \).

For the counterfactual images, we assume a causal graph satisfying the SCM (Structured Causal Model) condition in [15] and modeling the relation between images, attributes \( Y \) and sensitive attribute \( A \). Then, we conduct (1) abduction, (2) action, (3) prediction. The \( X_{A \leftarrow a' \leftarrow a} \) denotes the image generated by the intervention.

Then we define the counterfactually fair classifier \( f_\theta(\cdot)_i \) for the \( i \)-th attribute of \( y \) as:

\[
P(f_\theta(X_{A \leftarrow a})_i = y_i | X = x, A = a) = P(f_\theta(X_{A \leftarrow a'})_i = y_i | X = x, A = a).
\] (1)

Spontaneously the counterfactual fairness is measured by the counterfactual disparity (CDP) as follow:

\[
CDP_i(x,a) = |P(f_\theta(X_{A \leftarrow a})_i = y_i | X = x, A = a) - P(f_\theta(X_{A \leftarrow a'})_i = y_i | X = x, A = a)|,
\] (2)

where \( i = 1, \ldots, M \). The goal of this work is to simultaneously minimize the attribute prediction error \( err(\hat{y}, y) \) and the counterfactual disparity \( CDP_i(x,a) \) for each given face image.

4. Method

The proposed learning method consists of the following three procedures: (1) Discovering the causal relationships among the attributes to be used in a detailed causal graph, (2) learning the causal graph-based attribute translator for generating counterfactual images \( X_{A \leftarrow a'} \) from the given image, sensitive attribute, and causal relationship, and (3) learning the counterfactually fair facial attribute classifier with counterfactual regularization.

The overall framework for the proposed learning method is illustrated in Figure 1. Details of the figure are discussed below.
Figure 1. Illustration of the proposed framework. (a) The generation and inference process of the causal graph-based counterfactual attribute translator. Solid and dashed lines indicate the generation and inference process, respectively. The relationships between the sensitive attribute $A$ and facial attributes $Y$ are modeled in Bayesian Network (BN). The encoder $G_{enc}$ and $G_{dec}$ are modeled in autoencoder with adversarial loss. (b) The proposed learning framework. From the observed image $X$, two different types of counterfactual images are generated using the Causal Graph-based Counterfactual Attribute Translator. The total loss function consists of cross-entropy loss between $\hat{y}$ and $y$ and the counterfactual regularization that absolute difference between predictions of the generated facial images $\hat{y}_{A \leftarrow a}$ and $\hat{y}_{A \leftarrow a'}$. where $Y'$ is generated on the intervention of $A = a'$. $X_{A \leftarrow a} = G_{dec}(U, y', a')$ where $y'$ is a sample from intervened distribution of $Y$.

4.1. Discovering the Causal Relationship of Attributes

A face image is encoded into the following three variables: $A$, $Y$, and $U_{style}$. Here, $U_{style}$ is a latent variable independent of $Y$ and $A$ [48]. To generate counterfactual face images, it is necessary to discover the causal relationship among $Y$ and $A$.

In this paper, causal discovery based on a data-driven method is used to avoid additional biases from human perception. The causal relationships among $Y$ and $A$ are discovered by Graphical Lasso [49], and Greedy Equivalence Search (GES) [50] algorithms. The search space (the number of potential causal relationships among attributes) grows exponentially as the number of facial attributes increases. The Graphical Lasso provides sparse un-directed relationships between attributes, which helps to reduce the search space of causality discovery algorithms such as GES. The learned structure of the causal model is represented as an Adjacency matrix ($A_{adj}$).

Using the discovered causal graph structure, a Bayesian Network is trained to estimate the weights of causal relationships among attributes $Y$ and $A$. Then, we can have the causal model to address the dependency between attributes, which can be used for the prediction of counterfactual images.

4.2. Causal Graph-Based Attribute Translator for Counterfactual Images

The causal graph-based attribute translator is designed to generate multiple counterfactual facial images corresponding to the observed face image. It consists of two submodules:Encode for the abduction for $U_{style}$ in the proposed causal model, (2) attribute translator for the action, (3) decoder for the prediction. To do this, we use the convolutional neural network $G$, composed of the encoder $G_{enc}$ and decoder $G_{dec}$ connected to the casual graph-based sampler for attributes. The encoder $G_{enc}$ is for the abduction, and the decoder $G_{dec}$ combined with causal graph sampler is for the action and prediction. Figure 1a represents the generation and inference process of the Causal Graph-based Attribute Translator.
The image encoder \(G_{enc}\) for the it abduction takes the observed image \(x, y, A\) and extracts the latent representation \(U_{style}\). This is given below:

\[
U_{style} = G_{enc}(x, y, A),
\]

which is to model the posterior of \(U_{style}\).

In the attribute translator for the action, the counterfactual attribute vectors are sampled from the intervened attributes distribution. In sampling, we use the likelihood-weighted particle generation [51].

Then the image \(X\) is predicted by the decoder \(G_{dec}\) with the sampled attributes. The image decoder \(G_{dec}\) generates a realistic face image from the latent representation \(U_{style}\) and the sampled counterfactual attribute vector \(y_{A\to a'}\). This is given below:

\[
x_{A\to a'} = G_{dec}(U_{style}, y_{A\to a'}, a').
\]

To sum up, the neural network \(G\), referred to as causal graph-based attribute translator, generates a counterfactual image with respect to a given observed image \(x\) and counterfactual attribute \(y_{A\to a'}\):

\[
x_{A\to a'} = G(x, y_{A\to a'}, a'), \tag{3}
\]

and we consider two types of intervention distribution [52], i.e., \(X_{A\to a}\) and \(X_{A\to a'}\) because our aim is to construct the counterfactually fair classifier.

What remains is how to train the causal graph-based attribute translator. Two auxiliary classifiers are created to train the \(G\): (1) A discriminator \(D\) provides the supervision for generating realistic images, and (2) an auxiliary attribute classifier \(h\) forces the intervention image to have the sampled counterfactual attribute vector. The causal graph-based attribute translator is obtained from the training using the following losses:

\[
\begin{align*}
\mathcal{L}_{D, advo} &= -E_{(x, y)}[D(x)] + E_{(x, \tilde{y})}[D(G(x, \tilde{y}, \tilde{a}))], \\
\mathcal{L}_{G, advo} &= -E_{(x, y)}[D(G(x, \tilde{y}, \tilde{a})]], \\
\mathcal{L}_{h, cls} &= E_{(x, y)}[CE(h(G(x, \tilde{y}, \tilde{a}), \tilde{y})], \\
\mathcal{L}_{G, cls} &= E_{(x, \tilde{y})}[CE(h(G(x, \tilde{y}, \tilde{a}), \tilde{y})], \\
\mathcal{L}_{l1} &= ||G(x, y, a) - x||_1,
\end{align*}
\]

where \(CE\) is cross-entropy loss and \(\tilde{y}\) follows the mixture distribution of \(0.5Y_{A\to a} + 0.5Y_{A\to a'}\). Here, the \(\tilde{a}\) is tied with the \(\tilde{y}\) if the \(\tilde{y}\) is sampled from \(Y_{A\to a}\) then the value of \(\tilde{a}\) is \(a\), otherwise \(a'\). The optimization for \(G, D\), and \(h\) are formulated as below, with hyperparameters \(\lambda_{G, cls}, \lambda_{h, cls}, \lambda_{l1}\):

\[
\begin{align*}
G^* &= \arg \min \mathcal{L}_{G, advo} + \mathcal{L}_{l1} + \lambda_{G, cls} \mathcal{L}_{G, cls}, \\
D^*, h^* &= \arg \min \mathcal{L}_{D, advo} + \lambda_{h, cls} \mathcal{L}_{h, cls}.
\end{align*}
\]

The neural networks \(G, D\), and \(h\) are optimized, and the causal graph-based attribute translator is obtained by the optimum \(G^*\).

4.3. Counterfactually Fair Classifier

The causal graph-based attribute translator provides the generated counterfactual facial images corresponding to the given facial image. We propose the counterfactual regularization loss, shown below, for reducing the counterfactual disparity between predictions of factual and counterfactual images:
\[
L_{\text{cf-reg}}(x,a) = \frac{1}{M} \sum_{i=1}^{M} \left| P(\hat{y}_{A \leftarrow a, i} = 0 | X = x, A = a) - P(y_{A \leftarrow a', i} = 0 | X = x, A = a) \right|
\]

where \(\hat{y}_{A \leftarrow a, i} = f_\theta(X_{A \leftarrow a})\) and \(\hat{y}_{A \leftarrow a', i} = f_\theta(X_{A \leftarrow a'})\). Note that \(\sum_{i=0}^{3} P(\hat{y}_{A \leftarrow a, i} = y_i | X = x, A = a) = 1\). The facial attribute classifier \(f_\theta\) is a deterministic function of the input facial image, implemented as a convolutional neural network. And \(X_{A \leftarrow a'}\) is random based on the causal-graph. To estimate \(P(\hat{y}_{A \leftarrow a, i} = 0 | X = x, A = a)\), we take the strategy of sampling multiple images from two intervention distributions \(X_{A \leftarrow a}\) and \(X_{A \leftarrow a'}\) based on the causal graph. Note that

\[
x_{A \leftarrow a} = G^*(x, y_{A \leftarrow a, a}), \quad y_{A \leftarrow a} \sim \mathcal{B}N(Y|\text{do}(A = a)),
\]

\[
x_{A \leftarrow a'} = G^*(x, y_{A \leftarrow a', a'}), \quad y_{A \leftarrow a'} \sim \mathcal{B}N(Y|\text{do}(A = a')).
\]

The classification loss for facial attributes is as follow:

\[
L_{\text{cls}}(\hat{p}_i, y_i) = y_i \cdot \log(\hat{p}_i)
\]

where \(y_i, \hat{p}_i\) denote the \(i\)-the attribute of ground truth attribute vector \(y\) and its predicted probability, respectively. The counterfactual regularization \(L_{\text{cf-reg}}\) is added to the classification loss \(L_{\text{cls}}\). The overall loss function \((L_{\text{cf}})\) for the counterfactually fair classifier \(f_\theta\) is follows:

\[
L_{\text{cf}}(x, y, a) = \frac{1}{M} \sum_{i=1}^{M} L_{\text{cls}}(\hat{p}_i, y_i) + \lambda_{\text{cf}} L_{\text{cf-reg}}(x, a).
\]

During the training, we calculate the regularization loss of only those attributes whose intervention predictions \(\hat{y}_{A \leftarrow a}, \hat{y}_{A \leftarrow a'}\) (sampled from both intervention attribute distributions, \(\text{do}(A = a)\) and \(\text{do}(A = a')\)) are the same.

5. Experiments
5.1. Dataset

Large-scale CelebFace Attributes (CelebA) [2] is used for evaluating the proposed learning method for counterfactual fairness. The dataset consists of approximately 200,000 cropped and aligned facial images. Each facial image has 40 manually labeled binary attributes. In this work, the 29 attributes are selected to reduce the effect caused by extremely imbalanced label distribution; some attributes have too few observed samples in the training dataset to learn a causal controller. Gender attribute is selected as the sensitive attribute, and we conduct experiments around mitigating the bias caused by the different gender in each individual. Attributes are discarded when either the number of attributes samples is less than two samples for reliably predicting the attribute or there is an extreme imbalance in that the number of samples exceeds more than 100 times between the two demographic groups (male and female). The training, validating, and testing partition for experiments is the same as for the pre-defined set.

5.2. Experimental Details

For all experiments in this paper, the gender and the other 29 facial attributes were selected for the sensitive \(A\) and target attribute vector \(y\), respectively.

The causal graph-based attribute translator is trained to generate facial images with \((128 \times 128)\) width and height resolution. In this work, the architecture of the proposed attribute translator follows AttGAN [48] design, one of the widely used architectures for generating facial images with given attributes. To train the proposed attribute translator, Adam optimizer [53] is used with a \(2 \times 10^{-4}\) learning rate. The gradient penalty [54] is adopted for stable training of adversarial loss. The values of \(\lambda_{\text{G}, \text{cls}}, \lambda_{\alpha, \text{cls}}, \lambda_{1}\) are set to 100.0, 10.0, and 1.0, respectively along the AttGAN [48] to match the scale of different types of losses. The number of samples in the intervention distribution is set to 100 for each image observed for training, validation, and testing. The pre-trained ResNet-18 [55], etc. are adopted for stable training of adversarial loss.
one of the widely used convolutional neural networks in various computer vision tasks, is used as the backbone of the face attribute classifier. The framework is not limited to specific architecture design, which means the causal graph-based attribute translator and the counterfactually fair classifier could be replaced with a state-of-the-art conditional generator and facial attribute classifier.

Counterfactual augmentation is an algorithm that can be compared based on a common causal graph. Counterfactual augmentation is a method to address factual and counterfactual face images during the training of the face attribute classifier. During the training of the facial attribute classifier of both counterfactual augmentation and the proposed learning method, the causal graph-based facial attribute translator is used to generate factual and counterfactual facial image generation. In addition, “baseline” is an experiment evaluated using the counterfactual image generated by counterfactual augmentation of facial image translation to compare the counterfactual fairness performance of counterfactual augmentation and the proposed learning.

The classifier is trained using the union set of observed images and generated images using cross-entropy loss without counterfactual regularization. In the experimental results, the counterfactual augmentation is referred to as cf\_aug. The proposed counterfactually fair classifier is denoted by “cf\_reg” in the experimental results. The hyperparameter for counterfactual regularization loss, $\lambda_{cf}$, controls the effectiveness of the counterfactual regularizer. This parameter provides the controllability of the trade-off relationship between counterfactual parities and facial attribute classification accuracy. The range of $\lambda_{cf}$ in Section 5.5 is determined to examine the trade-off relationship between accuracy and counterfactual parities. All the experiments of this paper were conducted on a single Titan V, which has 12 GB of memory.

5.3. Causal Graph-Based Attribute Translator

5.3.1. Discovered Causal Structure

The discovered adjacency matrix representing causal structure is illustrated in Figure 2. The $y$-axis and $x$-axis indicate the cause and the effect, respectively. According to the discovered adjacency matrix, the gender attribute affects many other attributes directly or indirectly. The attributes which are influenced by gender are listed in Table 1.

In total, 14 of 29 attributes are directly affected, and 6 of the remaining attributes are indirectly affected. The number of edges is 104 which indicates that there are complex causal relationships among the facial attributes $y$ and the sensitive attribute $A$. For example, the attribute Eyeglasses is influenced by gender in the discovered causal structure, possibly because female celebrities tend to wear contact lenses instead of eyeglasses.

Table 1. Influenced attributes by gender in the CelebA.

| Sensitive Attribute | Male |
|---------------------|------|
| Directly Influenced Attributes | Arched_Eyebrows, Attractive, Bags_Under_Eyes, Bangs, Blond_Hair, Bushy_Eyebrows, Chubby, Double_Chin, Eyeglasses, Gray_Hair, No_Beard, Pointy_Nose, Wavy_Hair, Wearing_Hat |
| Indirectly Influenced Attributes | Oval_Face, Receding_Hairline, Black_Hair, Brown_Hair, Narrow_Eyes |
5.3.2. Sampled Counterfactual Attributes

One of the major motivations of this work is that the sensitive attribute has indirect influences on the facial attributes $y$. To visualize the concealed influence of gender, UMAP [56] is used. In Figure 3a, $y$ of each demographic group which shares the same sensitive attribute is easily distinguished from the other demographic group. On the other hand, when both $y$ and $y_a$ are represented in UMAP at the same time, the demographic groups cannot be distinguished in Figure 3b. Based on the UMAP visualizations, the counterfactual attribute vector $y_{A-a'}$ helps to narrow down the gap between the demographic groups.

5.3.3. Generated Counterfactual Images

The generated images with the intervention attribute distribution which corresponds to the observed image $x$ are illustrated in Figure 4. The pairs labeled $(x, y), (x_{A-a}, y_{A-a})$, and $(x_{A-a'}, y_{A-a'})$ are the pairs observed in the test dataset, pairs sampled from an intervention distribution with the same value of the sensitive attribute, and pairs sampled from a counterfactual distribution, respectively. The proposed causal-graph based attribute translator can generate realistic facial images from the intervention attribute and given image. In Figure 4, the attributes of translated images are matched to the intervention attribute vectors.
Figure 3. The UMAP visualization is applied for 29 binary facial attributes $y$ without the sensitive attribute (gender). (a) Visualization for observed $y$ in the training dataset. Facial attribute vectors of female ($y_A = 0$) and male ($y_A = 1$) are colored in orange and blue, respectively. (b) Visualization for observed $y$, and intervened $y_{A \leftarrow a'}$. The counterfactual attribute vectors of female and male are colored in green and dark red, respectively.

Figure 4. Observed and sampled images from intervention distribution on the gender. Nine of the twenty-nine target attributes and the sensitive attribute are visualized for $y$, $y_{A \leftarrow a}$, and $y_{A \leftarrow a'}$. The gray and red cells indicate observed facial attributes and their values switched due to the interventions, respectively.
5.4. The Effects of the Counterfactual Regularization

The expectation of the counterfactual disparity over the face images in test partition $E(CDP_i) = E_{x,a}(CDP_i(x,a))$ is used to measure counterfactual fairness for each attribute to focus the statistics of prediction over the images in test partition. The averaged counterfactual disparity of the baseline, cf_aug and proposed counterfactual regularization with different balancing hyperparameter $\lambda_{cf}$ are compared in Table 2.

Table 2. The comparison of the averaged counterfactual disparities for various algorithms on CelebA. The averaged counterfactual disparities are measured for $\lambda_{cf} = 0.01$, 0.05, and 0.1 in the columns cf_reg. The bold number represents the lowest value of averaged counterfactual disparity per attribute.

| Attribute            | Baseline | cf_aug | cf_reg |
|----------------------|----------|--------|--------|
|                      |          | 0.01   | 0.05   | 0.1    |
| Arched_Eyebrows      | 0.202    | 0.155  | 0.121  | 0.086  | 0.015  |
| Attractive           | 0.122    | 0.121  | 0.137  | 0.084  | 0.035  |
| Bags_Under_Eyes      | 0.111    | 0.126  | 0.024  | 0.043  | 0.008  |
| Bangs                | 0.045    | 0.026  | 0.028  | 0.031  | 0.005  |
| Big_Lips             | 0.116    | 0.092  | 0.037  | 0.043  | 0.023  |
| Big_Nose             | 0.153    | 0.116  | 0.064  | 0.090  | 0.039  |
| Black_Hair           | 0.055    | 0.041  | 0.011  | 0.060  | 0.004  |
| Blond_Hair           | 0.062    | 0.041  | 0.036  | 0.056  | 0.010  |
| Brown_Hair           | 0.036    | 0.041  | 0.072  | 0.040  | 0.024  |
| Bushy_Eyebrows       | 0.053    | 0.055  | 0.009  | 0.029  | 0.002  |
| Chubby               | 0.042    | 0.028  | 0.002  | 0.011  | 0.001  |
| Double_Chin          | 0.034    | 0.018  | 0.002  | 0.008  | 0.000  |
| Eyeglasses           | 0.030    | 0.032  | 0.003  | 0.018  | 0.003  |
| Gray_Hair            | 0.016    | 0.015  | 0.003  | 0.016  | 0.001  |
| High_Cheekbones      | 0.131    | 0.103  | 0.195  | 0.122  | 0.021  |
| Mouth_Slightly_Open  | 0.106    | 0.074  | 0.048  | 0.053  | 0.037  |
| Narrow_Eyes          | 0.045    | 0.024  | 0.020  | 0.028  | 0.007  |
| No_Beard             | 0.194    | 0.214  | 0.138  | 0.095  | 0.134  |
| Oval_Face            | 0.062    | 0.053  | 0.050  | 0.030  | 0.009  |
| Pale_Skin            | 0.022    | 0.024  | 0.000  | 0.005  | 0.000  |
| Pointy_Nose          | 0.088    | 0.096  | 0.053  | 0.052  | 0.010  |
| Receding_Hairline    | 0.036    | 0.039  | 0.011  | 0.015  | 0.003  |
| Smiling              | 0.108    | 0.092  | 0.060  | 0.066  | 0.064  |
| Straight_Hair        | 0.034    | 0.036  | 0.066  | 0.055  | 0.016  |
| Wavy_Hair            | 0.087    | 0.070  | 0.087  | 0.066  | 0.029  |
| Wearing_Earrings     | 0.121    | 0.077  | 0.009  | 0.064  | 0.009  |
| Wearing_Hat          | 0.029    | 0.032  | 0.005  | 0.014  | 0.002  |
| Wearing_Necklace     | 0.086    | 0.076  | 0.082  | 0.060  | 0.013  |
| Young                | 0.144    | 0.084  | 0.102  | 0.066  | 0.036  |
| **Average**          | 0.082    | 0.069  | 0.0501 | 0.048  | 0.019  |

The averaged counterfactual disparities for the baseline facial attribute classifier is measured to be approximately 0.082. The largest value of averaged counterfactual disparity is 0.202 for Arched_Eyebrows attributes. With the proposed counterfactual regularization, the trained counterfactually fair facial attribute classifier achieved 0.015 as the averaged counterfactual disparity for Arched_Eyebrows attribute. The averaged counterfactual disparities for over 27 attributes except for 2, reduces with the proposed counterfactual regularization. The lowest value of averaged counterfactual disparity is achieved with the cf_reg at $\lambda_{cf} = 0.1$. With the larger value of $\lambda_{cf}$, the measured averaged counterfactual disparities are decreased. Taking advantage of the fact that the counterfactual disparity $CDP_i(x,a)$ can be measured from each individual image, we compare the variance of $CDP_i(x,a)$. The meaning of the $\text{Var}(CDP_i(x,a))$ is the statistics for the degree of inconsistency in predicting $i$-th attribute by each individual. The averaged $\text{Var}(CDP_i(x,a))$ values over all attributes are 0.015, 0.014, and 0.004 for baseline, cf_aug, and cf_reg, respectively. Especially, the val-
ues of variances for attributes Eyeglasses, Smiling, Attractive are reduced remarkably, relative to $E(CDP_i)$. Note that the inference speed of the counterfactually fair classifier is the same as ResNet-18 [55] and there is no overhead during the inference because the attribute translator is used during the training procedure but not the inference procedure.

5.5. Trade-Off Relationship

The trade-off relationships between (1) averaged counterfactual fairness vs. averaged accuracy, (2) averaged counterfactual disparity vs. $\lambda_{cf}$, and (3) averaged accuracy vs. $\lambda_{cf}$ are plotted in Figure 5. The larger value of $\lambda_{cf}$ tends to possess less averaged counterfactual disparity, meaning that it is more fair, but the averaged accuracy is reduced. The counterfactual disparity and prediction accuracy have a linear relationship, and the decrease relative to the baseline (red dots in Figure 5) is sharper for the counterfactual fairness.

![Figure 5. Trade-off analysis of the proposed learning method for counterfactual fairness. Here, the averaged counterfactual disparity is defined as the aggregation over all $E(CDP_i)$ and the averaged prediction accuracy in the same manner. Each blue dot is obtained with a different value of $\lambda_{cf}$, and the red dot represents baseline performance.](image)

6. Conclusions and Discussion

This paper focuses on achieving counterfactual fairness for face attribute classification using a causal graph-based attribute translation. The proposed causal graph-based attribute translator consists of two submodules: (1) the Bayesian Network for modeling intervention attribute distribution and (2) an encoder–decoder framework for generating the intervention images from the given observed image. The counterfactual regularization, which reduces the counterfactual disparities, is proposed for multi-label classification. In the experimental results on the CelebA dataset, the intrinsic influences caused by the sensitive attribute are discussed, and the proposed causal graph-based attribute translator generates realistic intervened images from the observed face image through sampled intervened attribute vectors on the sensitive attribute. The comparison results show that the proposed learning method reduces both mean and variance of counterfactual disparity with a huge gap. The trade-off relationship between the accuracy and the counterfactual disparity with different hyperparameters is explored. Our work addresses fair facial attribute classification, and can be extended to various tasks such as identity verification and tagging systems.

Author Contributions: Conceptualization, S.K., G.K.; methodology, S.K. and G.K.; software, S.K.; validation, S.K.; formal analysis S.K., G.K. and C.D.Y., writing S.K., G.K. and C.D.Y.; supervision, G.K., and C.D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00951 Development of Uncertainty-Aware Agents Learning by Asking Questions and No.2022-0-00184 Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics).
Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kang, S.; Lee, D.; Yoo, C.D. Face attribute classification using attribute-aware correlation map and gated convolutional neural networks. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 4922–4926.

2. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 3730–3738.

3. Sun, Y.; Yu, J. Deep Facial Attribute Detection in the Wild: From General to Specific. In Proceedings of the 2018 British Machine Vision Conference, Newcastle, UK, 3–6 September 2018; p. 283.

4. Kumar, N.; Belhumeur, P.; Nayar, S. Facetracer: A search engine for large collections of images with faces. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 340–353.

5. Jang, Y.; Gunes, H.; Patras, I. Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild. Comput. Vis. Image Underst. 2019, 182, 17–29. [CrossRef]

6. Bhattacharj, B.; Bodur, R.; Kim, T.K. Auglabel: Exploiting word representations to augment labels for face attribute classification. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2308–2312.

7. Hand, E.M.; Chellappa, R. Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.

8. Minaee, S.; Luo, P.; Lin, Z.; Bowyer, K. Going deeper into face detection: A survey. arXiv 2021, arXiv:2103.14983.

9. Hardt, M.; Price, E.; Srebro, N. Equality of opportunity in supervised learning. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.

10. Zafar, M.B.; Valera, I.; Gomez Rodriguez, M.; Gummadi, K.P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International World Wide Web Conference, Perth, Australia, 3–7 April 2017; pp. 1171–1180.

11. Zafar, M.B.; Valera, I.; Rodriguez, M.G.; Gummadi, K.P. Fairness Constraints: Mechanisms for Fair Classification; PMLR: Fort Lauderdale, FL, USA, 2017; Volume 54, pp. 962–970.

12. Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; Zemel, R. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, USA, 8–10 January 2012; pp. 214–226.

13. Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; Huq. A. Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; pp. 797–806.

14. Grbic-Hlaca, N.; Zafar, M.B.; Gummadi, K.P.; Weller, A. The case for process fairness in learning: Feature selection for fair decision making. In Proceedings of the NIPS Symposium on Machine Learning and the Law, Barcelona, Spain, 5–10 December 2016; Volume 1, p. 2.

15. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4066–4076.

16. Wang, Y.; Kosinski, M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. J. Personal. Soc. Psychol. 2018, 114, 246. [CrossRef] [PubMed]

17. Taigman, Y.; Yang, M.; Ranzato, M.; Wolf, L. Deepface: Closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1701–1708.

18. Yu, H.X.; Zheng, W.S. Weakly supervised discriminative feature learning with state information for person identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5528–5538.

19. Shi, Y.; Jain, A.K. Probabilistic face embeddings. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6902–6911.

20. Wang, M.; Deng, W.; Hu, J.; Tao, X.; Huang, Y. Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.

21. Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, 23–24 February 2018; Friedler, S.A., Wilson, C., Eds.; PMLR: New York, NY, USA, 2018; Volume 81, pp. 77–91.
22. Kievit, R.; Frankenhuysen, W.E.; Waldorp, L.; Borsboom, D. Simpson’s paradox in psychological science: A practical guide. *Front. Psychol.* **2013**, *4*, 513. [CrossRef] [PubMed]

23. Baeza-Yates, R. Bias on the web. *Commun. ACM* **2018**, *61*, 54–61. [CrossRef]

24. Wang, T.; Wang, D. Why Amazon’s ratings might mislead you: The story of herding effects. *Big Data* **2014**, *2*, 196–204. [CrossRef] [PubMed]

25. Kärkkäinen, K.; Joo, J. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv* 2019, arXiv:1908.04913.

26. Kocaoglu, M.; Snyder, C.; Dimakis, A.G.; Vishwanath, S. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, 30 April–3 May 2018.

27. Hendricks, L.A.; Burns, K.; Saenko, K.; Darrell, T.; Rohrbach, A. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, Munich, Germany, 8–14 September 2018; pp. 793–811.

28. Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.W.; Ordonez, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, 27 October–2 November 2019; pp. 3310–3319.

29. Alvi, M.; Zisserman, A.; Nelläker, C. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the 2018 European Conference on Computer Vision* (ECCV), Munich, Germany, 8–14 September 2018.

30. Ryu, H.J.; Adam, H.; Mitchell, M. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv* 2017, arXiv:1712.00193.

31. Xu, D.; Yuan, S.; Zhang, L.; Wu, X. Fairgan: Fairness-aware generative adversarial networks. In *Proceedings of the 2018 IEEE International Conference on Big Data* (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 570–575.

32. Wang, Z.; Qinami, K.; Karakozis, I.C.; Genova, K.; Nair, P.; Hata, K.; Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 14–19 June 2020; pp. 8919–8928.

33. Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; Chang, K.W. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, 7–11 September 2017; pp. 2979–2989. doi: 10.18653/v1/D17-1323. [CrossRef]

34. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4066–4076.

35. Kusner, M.; Russell, C.; Loftus, J.; Silva, R. Making decisions that reduce discriminatory impacts. In *Proceedings of the International Conference on Machine Learning*, Long Beach, CA, USA, 10–15 June 2019; pp. 3591–3600.

36. Zhang, J.; Bareinboim, E. Equality of opportunity in classification: A causal approach. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2018*, Montreal, QC, Canada, 3–8 December 2018; pp. 3671–3681.

37. Chiappa, S.; Isaac, W.S. A causal Bayesian networks viewpoint on fairness. In *Proceedings of the IFIP International Summer School on Privacy and Identity Management*, Vienna, Austria, 20–24 August 2018; pp. 3–20.

38. Chiappa, S. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7801–7808.

39. Joo, J.; Kärkkäinen, K. Gender Slopes: Counterfactual Fairness for Computer Vision Models by Attribute Manipulation. *arXiv* 2020, arXiv:2005.10430.

40. Denton, E.; Hutchinson, B.; Mitchell, M.; Gebru, T. Detecting bias with generative counterfactual face attribute augmentation. *arXiv* 2019, arXiv:1906.06439.

41. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems 2014*, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.

42. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, 30 April–3 May 2018.

43. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA, 6–9 May 2019.

44. Kocaoglu, M.; Snyder, C.; Dimakis, A.G.; Vishwanath, S. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, 30 April–3 May 2018.

45. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 14–19 June 2020; pp. 8188–8197.

46. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the 2017 IEEE International Conference on Computer Vision* (ICCV), Venice, Italy, 22–29 October 2017.
49. Friedman, J.; Hastie, T.; Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **2008**, *9*, 432–441. [CrossRef] [PubMed]

50. Hauser, A.; Bühlmann, P. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **2012**, *13*, 2409–2464.

51. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.

52. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.

53. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference for Learning Representations, San Diego, CA, USA, 7–9 May 2015.

54. Thanh-Tung, H.; Tran, T.; Venkatesh, S. Improving Generalization and Stability of Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.

55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

56. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.