Do Counterfactually Fair Image Classifiers Satisfy Group Fairness? – A Theoretical and Empirical Study

Sangwon Jung[1*] **Sumin Yu**[1*] **Sanghyuk Chun**[2†] **Taesup Moon**[1,3†]

[1] Department of Electrical and Computer Engineering, Seoul National University
[2] NAVER AI Lab [3] ASRI/INMC/IPAI/AIIS, Seoul National University

Abstract

The notion of algorithmic fairness has been actively explored from various aspects of fairness, such as counterfactual fairness (CF) and group fairness (GF). However, the exact relationship between CF and GF remains to be unclear, especially in image classification tasks; the reason is because we often cannot collect counterfactual samples regarding a sensitive attribute, essential for evaluating CF, from the existing images (*e.g.*, a photo of the same person but with different secondary sex characteristics). In this paper, we construct new image datasets for evaluating CF by using a high-quality image editing method and carefully labeling with human annotators. Our datasets, CelebA-CF and LFW-CF, build upon the popular image GF benchmarks; hence, we can evaluate CF and GF simultaneously. We empirically observe that CF does not imply GF in image classification, whereas previous studies on tabular datasets observed the opposite. We theoretically show that it could be due to the existence of a latent attribute G that is correlated with, but not caused by, the sensitive attribute (*e.g.*, secondary sex characteristics are highly correlated with hair length). From this observation, we propose a simple baseline, Counterfactual Knowledge Distillation (CKD), to mitigate such correlation with the sensitive attributes. Extensive experimental results on CelebA-CF and LFW-CF demonstrate that CF-achieving models satisfy GF if we successfully reduce the reliance on G (*e.g.*, using CKD).

1 Introduction

As machine learning algorithms are deployed in societal computer vision applications such as facial recognition [39] and job interview [29], concerns have grown regarding their potential to discriminate against certain individuals and groups. For instance, a face recognition system might exhibit disparate accuracies across different demographic groups [3], while a job interview algorithm could be biased based on protective attributes even for the same interviewee [11]. Consequently, *algorithmic fairness* in image classifiers has gained significant attention in academic and industrial research communities.

While conceptually apparent, determining a concrete notion of fairness is challenging, leading to the proposal of several different fairness notions. One prevalent notion is *counterfactual fairness* (CF) [23] which seeks consistent predictions when only a sensitive attribute is intervened. Another important notion is *group fairness* (GF) [43] that aims to treat different demographic groups equally to prevent one group unfairly disadvantaged compared to another. Many researchers have focused on developing separate algorithms to achieve each notion, while understanding the exact relationship between CF and GF is yet under-explored; *e.g.,* some recent work [1, 35] showed that a model achieving CF can meet several GF notions *only* under specific conditions of Structural Causal Models.

[*] Equal contribution.
[†] Co-corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

Furthermore, previous studies on the relationship between CF and GF have not considered the setting of image classification due to the absence of *evaluation* datasets with counterfactual images, in which only the sensitive attribute is altered from the original images while other attributes not caused by the sensitive attribute remain unchanged — a data nearly impossible to collect in the real world. There have been several works generating counterfactual images using generative models [4, 19, 26, 32, 44, 34, 5], but they have only focused on utilizing generated counterfactual samples for training rather than evaluation. Moreover, these methods often suffer from low-quality counterfactual images generated based on VAE [21] or GAN [9]. One notable exception is Liang et al. [24], which offers an evaluation dataset including counterfactual images. However, their images are all synthetic; thus, it is still insufficient to evaluate CF due to distribution shifts from real-world images.

In this paper, we construct CF benchmarks for image classification tasks using high-performing diffusion model-based generative models. Our datasets build upon popular facial benchmark datasets used for evaluating GF, CelebA and LFW, by altering the sensitive attribute with pre-trained Instruct-Pix2Pix (IP2P) [2]. We then carefully curate the edited samples by human annotators and verify the reliability of our datasets as counterfactual samples from additional annotators. Note that our datasets, CelebA-Counterfactual Face (CelebA-CF) and LFW-Counterfactual Face (LFW-CF), share the same test samples as the original GF benchmarks, enabling the evaluation of both GF and CF.

Using our datasets, we conduct a primitive study on the relationship between CF and GF in image classification, *e.g.*, test whether CF implies GF for image classifiers. To that end, we train CF-aware methods [36, 7] and evaluate them with our datasets using both CF and GF metrics. From the result, we observe that they achieve CF but fail to satisfy GF, contrary to previous findings that CF can imply GF [1, 35]. We attribute this failure to Structural Causal Models (SCMs) of image generation. Specifically, for an image SCM, a latent attribute G is more likely to exist, which could be correlated with, but not caused by, the sensitive attributes. For example, in the case where the sensitive attribute is the sex of a person in an image, secondary sex characteristics such as beard and hairline are highly correlated with hair length, but it does not mean that such characteristics cause the length of hair. In this scenario, if a model achieving CF relies on the attribute G (*e.g.*, hair length) on its prediction, it could more severely violate GF in the worst case. Therefore, if we can reduce the dependency on G of a CF-aware model, we may achieve both CF and GF. Empirically, we find that a model trained with vanilla cross-entropy loss is more robust to G than a model trained with a CF-aware method. Motivated by this, we propose a simple baseline, named Counterfactual Knowledge Distillation (CKD), which distills the robustness to G during the original CF-aware optimization. Finally, our extensive experiments using CelebA-CF and LFW-CF demonstrate that CF-achieving models satisfy GF when reducing the reliance on G (*e.g.*, using CKD).

In summary, our contributions are three-fold. Firstly, we construct two new image classification benchmarks for measuring CF, CelebA-CF and LFW-CF. Secondly, using these datasets, we observe the disparity between CF and GF in image classifiers and provide a theoretical rationale; a counterfactually fair classifier may not necessarily achieve GF when an additional latent attribute that is correlated with the sensitive attribute exists. Finally, we propose a simple baseline, CKD, to reduce the sensitivity to such latent attributes of a model, resulting in achieving CF and GF simultaneously.

2 Constructing high-quality counterfactual images

The degree of counterfactual fairness (CF) can be measured by the prediction consistency between an original sample and its corresponding counterfactual (CTF) sample. For a given sample and a sensitive attribute, a CTF sample is defined as the one of which the sensitive attribute is altered while all the other attributes not caused by the sensitive attribute remain the same. However, acquiring a CTF sample for an image is challenging. For example, if the sex of a person in an image is the sensitive attribute, obtaining a CTF sample requires changing the secondary sex characteristics of the person such as beard or hairline, while preserving their identity and the other attributes, which is impossible in practice. One possible alternative is to generate a virtual face by altering such secondary sex characteristics of the given identity using a high-quality image editing method.

Several previous approaches [19, 34, 41, 44, 22] have attempted to generate CTF images by VAE or GAN-based editing methods. However, they have struggled with low image quality or unintended modifications to non-sensitive attributes, rendering them unreliable for evaluating CF. To address such issues, we employ IP2P [2], an advanced diffusion model-based image editing method. Notably,

Figure 1: **CelebA-CF examples**. The counterfactual (CTF) images regarding the "sex attribute" are shown.

IP2P can generate high-quality CTF samples by simply adjusting the text instructions without any model retraining.

As the first step, we edit the test images of two popular facial image datasets, CelebA [25] and LFW [13]. We choose the "sex" of a person in an image as the sensitive attribute[3] and edit the sex-related visual characteristics of facial images using text prompts. We generated 720 CelebA CTF image pairs and 632 LFW CTF image pairs, where the images are selected to be balanced across groups for both target and sensitive labels. Here, we treat "blond hair" and "smiling" as the target labels for CelebA and LFW, respectively. Namely, for example, the CelebA CTF image pairs have a balanced group of <female, non-blond hair>, . . . , and so on. Figure 1 and A.1 show examples of generated CTF images together with the originals. Hyperparameter settings are reported in Appendix C.1. Note that while we adopt the "sex" attribute, our generation process is attribute-agnostic (*e.g.*, age or skin color can be also used in place of sex) as illustrated in Figure A.2.

Image filtering. Despite the high quality of IP2P, low-quality CTF images can still be generated. To address this, we employed five human annotators to filter the images, *i.e.*, each image pair was annotated as either "correct" or "incorrect". To ensure objective and precise annotation criteria, we created guidelines as follows. Initially, we compiled a list of 20 masculine and feminine visual features using GPT-4o and with guidance from experts specialized in fairness, selected nine key facial attributes representing sex-related visual characteristics: facial hair, Adam's apple, skin texture, jawline, chin shape, brow ridge, cheekbone prominence, lip fullness, and hairline. These attributes were used to establish the criteria for evaluating correct CTF samples. One notable issue is that most of the feminine-like images in CelebA and LFW datasets include makeups (for instance, many female celebrities in the CelebA dataset appear to be wearing makeup) and the IP2P model is biased towards removing makeup when altering feminine features. To prevent images from being filtered out solely due to changes in makeup, we additionally included makeup in the set of key attributes, even though it is not a sex characteristic. Finally, the guidelines were created based on these ten attributes, providing some criteria for correct CTF samples, such as whether the change of some of the ten attributes was accurate and whether other facial characteristics remained consistent with the original image. Using these guidelines, we filtered out pairs receiving two or fewer "correct" votes, resulting in 230 and 144 images for CelebA and LFW, respectively. More details about the human annotating interface are in Appendix C.2, and additional information on the newly created dataset can be found in Appendix C.4.

Reliability check. We further verify the quality of our datasets by additional five human annotators, distinct from those participated in the filtering process. Those annotators evaluate only the images that remained after the filtering, based on two criteria: (1) whether the sensitive attribute was correctly changed and (2) whether the other non-sensitive attributes were preserved. The annotators evaluated the images for the sensitive attribute, "sex" and three non-sensitive attributes, "blond hair", "gray hair", and "smiling"; we chose these three because other attributes can be subjective (*e.g.*, "big nose") [40] or had already been filtered (*e.g.*, "wearing hat"). Details of the annotating interface provided to

Table 1: **Human evaluation of the reliability of our datasets**. Accuracies of the correctly altered sensitive attributes and well-preserved non-sensitive attributes are shown.

	Sensitive	Non-Sensitive
CelebA-CF	96.52	95.98
LFW-CF	98.61	93.75

[3]The two datasets use the terms "gender" for indicating their sensitive attributes. However, using such terminology can present some ethical concerns because they can suggest meanings linked to social identities. Thus, we have decided to use the term "sex" instead, which more accurately refers to biological characteristics.

the five annotators are in Appendix C.2. Based on the majority vote, we compute the percentage of CTF samples which met each of the two criteria, *i.e.*, the accuracies for whether the sensitive and non-sensitive attributes are correctly altered and preserved. Table 1 displays the values for CelebA-CF and LFW-CF. The non-sensitive accuracy is averaged across three non-sensitive attributes. The results demonstrate that our CTF samples almost meet the two CTF criteria, suggesting that our datasets can be reliably utilized to evaluate CF.

Ethical considerations. In our study, we use the term "sex", not "gender", to represent the sensitive attribute with biological traits, because terms such as "gender" might imply associations with social identities, potentially raising some ethical issues. We also specifically choose ten perceived facial attributes as the visual features representing the biological sex in facial images. We believe that these considerations help alleviate various normative harms that arise from dichotomizing gender, which refers to social identity. However, despite our efforts, the sex-related visual characteristics are complex and intertwined, making it challenging to fully represent with a binary label. Thus, we urge practitioners to use our datasets with these considerations in mind.

3 Primitive study on the relationship between CF and GF

3.1 Experimental setup

We consider the image classification task where each data sample consists of an input image X, a class attribute $Y \in \mathcal{Y} = \{0, \cdots, |\mathcal{Y}| - 1\}$ and a sensitive attribute $A \in \{0, 1\}$, *e.g.*, sex.

Metrics. We measure three metrics for CF, GF, and classification accuracy. Firstly, we describe the metric for CF. A classifier satisfies CF when the predictions for the original sample and its counterfactual (CTF) sample are the same for every sample x and sensitive attribute a, *i.e.*, $P(\widehat{Y} = y|X = x, A = a) = P(\widehat{Y}_{A \leftarrow a'} = y|X = x, A = a)$, where $\widehat{Y}_{A \leftarrow a'}$ represents the prediction for a counterfactual sample intervened on A with a' (*e.g.*, changing female to male). We quantify the degree of violence with respect to CF using counterfactual disparity (CD):

$$\text{Counterfactual Disparity } \textbf{(CD)} \triangleq \mathbb{E}_{x,a}\left[P\left(\mathbb{1}\{\widehat{Y}_{A \leftarrow a'} \neq \widehat{Y}\}|x, a\right)\right]. \tag{1}$$

Secondly, we adopt equalized odds (EO) as our notion for GF. If a predictor $\widehat{Y}$ and the sensitive attribute A are conditionally independent given the true class attribute Y, the predictor satisfies EO; namely, EO holds when $P(\widehat{Y} = y'|A = 0, Y = y) = P(\widehat{Y} = y'|A = 1, Y = y)$. From the definition, we can capture the degree of violence with respect to GF with the disparity of EO (DEO):

$$\text{Disparity of EO } \textbf{(DEO)} \triangleq \max_{y,y' \in \mathcal{Y}} \left|P(\widehat{Y} = y'|A = 0, Y = y) - P(\widehat{Y} = y'|A = 1, Y = y)\right|. \tag{2}$$

We note that we empirically compute CD and DEO, defined in Equation (1) and (2), using our benchmark datasets and the original test datasets of CelebA and LFW, respectively. Additionally, Pinto et al. [33] propose several other metrics to evaluate CF, and accordingly, we conducted an additional evaluation based on these metrics, with results provided in Appendix G.5.

Baseline methods. We evaluate a model trained with the vanilla cross-entropy loss (denoted as "Scratch") and two CF-aware training methods, Scratch+aug and counterfactual pairing (CP). Scratch+aug is a Scratch method using an augmented training dataset with counterfactual samples [7], and CP [36] adopts a regularization term that promotes pairs of original and its CTF sample to obtain the same prediction (see Equation (4) for the formal definition). Note both methods need counterfactual samples for training, and hence, we use the samples generated via IP2P with the same prompts used in Section 2 without any filtering process to obtain results for them. For a comprehensive study, we additionally evaluate two individual fairness-aware methods, SenSeI [42] and LASSI [31], of which goals are analogous to CF in aiming to make a model robust to perturbation of the sensitive attribute. More details are described in Appendix D.3.

Model selection. Due to the accuracy-fairness trade-off [6], appropriate model selection is important for fair evaluation. We explore varying hyperparameters and select the best model that shows the lowest CD (Equation (1)) for the held-out validation set while achieving a lower bound of the accuracy[4].

[4]Considering the accuracy degradation of fair-training methods, we set the bound as 98% of the accuracy of Scratch, *i.e.*, if Scratch achieves 95.0% accuracy, then we only consider models with more than 93.1% accuracy.

Table 2: **CF does not always imply GF on image classification**. We report CD (Equation (1)) and DEO (Equation (2)) for measuring Counterfactual Fairness (CF) and Group Fairness (GF), respectively. Accuracy and DEO are measured on the original test datasets (CelebA and LFW) and CD is evaluated on the newly constructed datasets, CelebA-CF and LFW-CF, described in Section 2. If a model shows an inferior metric value than Scratch, the number is highlighted in **red**.

Method	CelebA (and CelebA-CF)			LFW (and LFW-CF)		
	Acc ↑	CD ↓	DEO ↓	Acc ↑	CD ↓	DEO ↓
Scratch	95.53	10.26	47.10	90.85	18.06	7.66
Scratch+aug [7]	95.41	4.65	44.71	90.34	12.15	**7.86**
CP [36]	94.10	2.53	**51.01**	89.77	9.20	**8.74**
SenSeI [42]	95.33	8.00	**52.32**	87.75	16.09	**9.23**
LASSI [31]	91.07	9.69	31.79	-	-	-

3.2 Performance comparison

Table 2 shows accuracy, CD, and DEO for Scratch and four baseline methods. Note that we omit the result of LASSI on LFW because the number of samples in LFW is not enough to train the Glow model [20], which is the main component of LASSI. From the table, CF-aware and individual fairness-aware methods are mostly effective in mitigating CD, when compared to Scratch. However, it does not necessarily lead to improvements in DEO. Especially, while CP significantly improves CD for both datasets, it exacerbates DEO compared to Scratch. Namely, contrary to the previous studies [1, 35] showing that CF implies GF on tabular datasets, our observation shows that CF does not always imply GF on image datasets. In the following section, we theoretically investigate why the previous observations may not hold on images.

4 Theoritical analysis on the relationship between CF and EO for images

4.1 Structural Causal Model (SCM) for images

Structural Causal Models (SCMs) are represented as directed acyclic graphs satisfying the conditions specified in [30]. In these models, nodes and edges indicate variables and their causal relationships within the data-generating process. As studied in previous works [4, 22], the nodes of an SCM for image can be categorized into three parts. As shown in Figure 2, the blue, gray and yellow nodes indicate latent attributes, *e.g.*, Y or A, components of the image influenced by these attributes, *e.g.*, X_Y or X_A, and the whole image X, respectively. Taking an SCM for facial images as an example, we can interpret these nodes as follows: latent attributes such as hair color or sex, facial components like the hair or an Adam's apple in a facial image, and an entire face. Note that the blue region in the figure describes that various causal relationships among latent attributes can exist [5]. Furthermore, although an image SCM may contain additional latent attributes, we simplify our focus to only include the class and sensitive attribute, Y and A, and a third-party attribute, G, which may correlate with the sensitive attribute A.

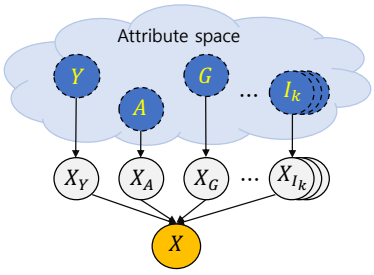

Figure 2: **Image SCM**. Blue, gray, and yellow circles represent latent attributes, components of an image and a whole image, respectively. Directed edges indicate a causal relationship from the source to the target. The blue region indicates that there can be any direction of edges between blue nodes.

4.2 Theoretical analysis

According to the Markov assumption of SCM [30], if there are no unblocked paths between two variables in an SCM (*i.e.*, they are d-separated), the variables are statistically independent. Utilizing this property, Anthis and Veitch [1] demonstrated that CF implies several GF notions, including Equalized Odds (EO) [10], under the specific condition on SCMs such as no *backdoor* path from the

[5] We assume no edge or unblocked path from A to Y; otherwise, all counterfactually fair models based on that SCM would produce random predictions with respect to X_Y.

sensitive attribute A to the image X exists (Theorem 2 of [1]). Moreover, the authors empirically show that these conditions would hold on some tabular datasets.

However, we argue that these conditions would not hold for image datasets due to a fundamental difference in what sensitive attributes represent in an image. Specifically, tabular datasets typically consist of recorded information by subjects, where sensitive attributes such as sex or race usually represent immutable genetic information; hence they are not caused by other attributes and cause all attributes correlated with the sensitive attributes. In contrast, sensitive attributes in image datasets indicate visual characteristics that can change and be influenced by some other attributes, such as the attribute G. For example, in a facial image dataset, attributes like hair length or accessories might be highly correlated with, but not caused by, secondary sex characteristics such as beard. Namely, a backdoor path from the sensitive attribute X through the attribute G could exist, thereby breaking the connection between CF and GF discovered in previous studies.

Our theoretical result specifies the relationship between CF and GF (especially for EO) with G:

Theorem 4.1. *Assume a latent attribute G in Figure 2 is a non-descendant variable of A and connected to A through an unblocked path. Then, the following inequality holds for a counterfactually fair classifier $\boldsymbol{\theta}$ and any pairs of y and y':*

$$\left|P(\widehat{Y}=y'|A=0,Y=y)-P(\widehat{Y}=y'|A=1,Y=y)\right| \leq \sum_{X_Y} P(X_Y|Y=y) \max_{X_G,X'_G} d_{\boldsymbol{\theta},X_Y}(X_G,X'_G), \tag{3}$$

in which $d_{\boldsymbol{\theta},X_Y}(X_G,X'_G)=\left|P(\widehat{Y}=y'|X_Y,X_G)-P(\widehat{Y}=y'|X_Y,X'_G)\right|$ and $\hat{Y}$ is the prediction of the model $\boldsymbol{\theta}$. The equality holds when $d_{\boldsymbol{\theta},X_Y}=0$ always regardless of X_Y.

The proof of the theorem is in Appendix A. Note that when we take the maximum over (y,y') on both sides of the inequality in Theorem 4.1, the left-hand side of the inequality becomes identical to DEO (Equation (2)). Therefore, the theorem implies that DEO is upper bounded by the maximum of $d_{\boldsymbol{\theta},X_Y}(X_G,X'_G)$ (in which the maximum is over X_G, X'_G, y, y'), which measures the sensitivity of the model with respect to G. In other words, the theorem shows that when a counterfactually fair model is sensitive to X_G (*i.e.*, when $\max d_{\boldsymbol{\theta},X_Y}(X_G,X'_G)$ is large), the model may result in having high DEO in the worst-case.

Theorem 4.1 elucidates why CF-aware methods in Table 2 often fail to mitigate DEO despite significant improvements in CD. Namely, if the attribute G assumed in Theorem 4.1 exists on CelebA and LFW, DEO for the classifiers trained by CF-aware methods can worsen depending on their robustness to G. This will be empirically demonstrated using "hair length" as G in Section 5.2, together with the results using a controllable synthetic dataset. Furthermore, Theorem 4.1 suggests that we can re-establish the relationship between two notions by making counterfactually fair classifiers non-sensitive to G. In the following section, we introduce a method to promote a classifier not to depend on G while achieving CF.

5 Empirical analyses on the effect of G to CF and GF

5.1 Counterfactual Knowledge Distillation (CKD)

Motivated by Theorem 4.1, we propose a baseline fair-training method to achieve both CF and GF. Conceptually, if we can reduce the dependency between the latent attribute G described in Theorem 4.1 and the prediction of a CF-aware trained model, we can expect that the model will achieve CF and GF simultaneously. Therefore, we improve the CF-aware method, CP [36] (best-performing in Table 2), such that the dependency to the attribute G is reduced. We first describe the CP regularization (which is used along with the cross-entropy loss) for given counterfactual samples $\mathcal{D}'=\{x_{i,A\leftarrow a'_i}\}_{i=1}^N$ corresponding to the original training dataset $\mathcal{D}$:

$$\mathcal{L}_{\text{CP}}(\theta,\mathcal{D}\cup\mathcal{D}') := \frac{1}{N}\sum_{i=1}^{N}\|f(\boldsymbol{\theta},x_i)-f(\boldsymbol{\theta},x_{i,A\leftarrow a'_i})\|_2^2, \tag{4}$$

in which $f(\boldsymbol{\theta},x)$ is a representation vector of input x produced by a classifier $\boldsymbol{\theta}$, such as logit or feature vector. Note that the images x and $x_{A\leftarrow a'}$ differ only in their components corresponding to

the sensitive attribute A and the attributes caused by the sensitive attribute A. Hence, although the CP regularization works well for achieving CF, it does not ensure the model does not rely on the attribute G, potentially leading to worse DEO as argued in the previous section.

Recent studies [16, 38, 45] have shown that the robustness of a teacher model can be transferred into a student model through knowledge distillation (KD) [12]. To that end, we first assume a teacher model that is robust to the attribute G is available. Then, our idea is to apply both KD and CP regularization to train our student model, which leads to a simple yet effective approach, dubbed as Counterfactual Knowledge Distillation (CKD). Specifically, CKD employs averaged representation vectors of original and counterfactual samples extracted by the teacher model $\boldsymbol{\theta}^T$ as target vectors. Then, representation vectors of both samples from the student model $\boldsymbol{\theta}$ are enforced to follow the target vectors. Namely, the distillation term of CKD is defined as follows:

$$\mathcal{L}_{CKD}(\boldsymbol{\theta}, \mathcal{D} \cup \mathcal{D}') := \frac{1}{2N} \sum_{i}^{N} \Big(\|f(\boldsymbol{\theta}, x_i) - f_i^T\|_2^2 + \|f(\boldsymbol{\theta}, x_{i,A \leftarrow a_i'}) - f_i^T\|_2^2 \Big),$$

$$\text{in which } \; f_i^T = \frac{1}{2}\big(f(\boldsymbol{\theta}^T, x_i) + f(\boldsymbol{\theta}^T, x_{i,A \leftarrow a_i'})\big) \; \text{ is the target vector for the } i\text{-th pair.} \tag{5}$$

Note that our distillation terms have both effects of KD and CP by promoting both representations of original and counterfactual samples to be aligned with the target vectors f_i^T produced by the teacher model. Therefore, based on Theorem 4.1, we can deduce that the CKD regularization encourages the model to achieve both CF (by the CP effect) and EO-based GF (by the KD effect that distills the robustness of the teacher with respect to the attribute G). In addition, we optionally incorporate CP regularization into our objective to further mitigate CD. The final objective of our method (which we again dub as CKD for brevity) is as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{\text{CE}}(\boldsymbol{\theta}, \mathcal{D}) + \mu \mathcal{L}_{\text{CKD}}(\boldsymbol{\theta}, \mathcal{D} \cup \mathcal{D}') + \lambda \mathcal{L}_{\text{CP}}(\boldsymbol{\theta}, \mathcal{D} \cup \mathcal{D}'), \tag{6}$$

in which μ and λ are controllable hyperparameters for the CKD and CP regularization, respectively.

While we assumed above the availability of a teacher model that is robust to the attribute G, obtaining such a model could be challenging in practice. Empirically, we observe that vanilla-trained models (referred to as "Scratch" models) less depend on the attribute G than CP-trained ones (see Figure 3 and Table 3 for more details). We presume that this is because the attributes A and G behave as "shortcut" features for classifying the class attribute Y, *i.e.*, they are easy-to-learn discriminatory features. As observed by Scimeca et al. [37], making a model blind to a certain shortcut feature causes it to rely more heavily on the other shortcut features. In our case, CP-trained models are trained to be invariant to the sensitive attribute A, resulting in a greater dependence on the attribute G compared to the Scratch models. Thus, unless otherwise specified, we will assume the vanilla-trained model is relatively robust to the attribute G since it would mostly rely on the sensitive attribute A, hence, we use it as the teacher model.

5.2 Impact of robustness to G on CF and GF

We empirically validate our theoretical result and CKD on both a newly introduced synthetic dataset (CIFAR-10B) and a real-world dataset (CelebA) by analyzing CF, GF, and the robustness with respect to the attribute G described in Theorem 4.1. We thus introduce a new metric for the robustness to the attribute G, the rate of flipped predictions (**RFP**) :

$$\text{RFP} \triangleq \mathbb{E}_{x,x'}\big[P\big(\mathbb{1}\{\widehat{Y} \neq \widehat{Y}'\}|x, x'\big)\big]. \tag{7}$$

in which x is an original image, x' is its corresponding image with the attribute G flipped. $\widehat{Y}$ and $\widehat{Y}'$ refer to the predicted label by the trained model given x and x', respectively. RFP quantifies the amount of flipped predictions when the attribute G is altered. For example, if a model shows the same prediction after changing the attribute G, its RFP becomes 0%.

CIFAR-10B, a controllable synthetic dataset. We construct the CIFAR-10B dataset, where we can perfectly control the degree of bias with respect to the attribute G while the target label is biased towards the sensitive attribute A. We make binary class labels from the 10 classes of CIFAR-10 (0-4 and 5-9 classes). We set the attributes A and G in Theorem 4.1 with the presence of Gaussian and Contrast noise, respectively. We also set a fixed ratio of 0.8 and a controllable ratio α, which represent

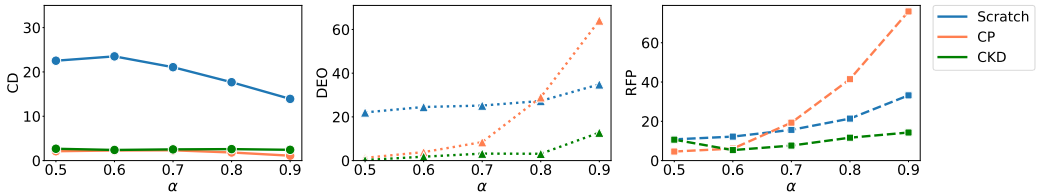

Figure 3: **Impact of the correlation of G and A.** α indicates how A and G are correlated on CIFAR-10B.

skewnesses among (Y, A) and (A, G), respectively; the former ratio is the spurious correlation between Y and A, and the latter one is the correlation between A and G. We then construct the CIFAR-10B dataset by randomly injecting Gaussian or Contrast noise to each CIFAR-10 image at given ratios, as illustrated in Figure E.2. Unless otherwise noted, we set α as 0.8.

We train models with Scratch, CP, and CKD on CIFAR-10B by adjusting α from 0.5 (*i.e.*, A and G are decorrelated) to 0.9 at intervals of 0.1. Figure 3 shows CD, DEO, and RFP metric values for each method. The figures indicate that while CP and our CKD consistently achieve CF, CP fails to meet GF as α increases, potentially due to higher RFP. Furthermore, RFP of Scratch is lower than that of CP when α is greater than 0.7. This empirically justifies the use of vanilla-trained models as teacher models robust to G. By using these teacher models, CKD significantly improves DEO regardless of the value of α by maintaining the robustness to G, *i.e.*, low RFP, supporting the result of Theorem 4.1.

Impact of G manipulation on CelebA. We assume "hair length" as G for facial image datasets, *e.g.*, CelebA because the hair length G can be highly correlated with, but not caused by, the sex A. To compute RFP for the hair length attribute, we manipulate the hair length of CelebA test images using SDEdit [28]. More details and generated examples can be found in Appendix E.1. Using the hair length-edited images, we report RFP in Table 3, together with DEO and CD. The results share the same trend as the CIFAR-10B results, *i.e.*, CP shows worse DEO and RFP than Scratch but better CD, whereas CKD shows the best DEO and RFP, despite a slight increase of CD.

Table 3: **Impact of G on CelebA.** We assume "hair length" as G and manipulate the hair length of test images. CD, DEO, and RFP are measured on CelebA-CF, CelebA, and hair-edited CelebA, respectively.

Method	CelebA CD ↓	CelebA DEO ↓	CelebA RFP ↓
Scratch	10.26	47.10	15.27
CP	**2.53**	51.01	20.37
CKD	4.44	**13.23**	**10.85**

5.3 Impact of the robustness to G of the teacher model on CKD

Our CKD requires a robust teacher model with respect to the attribute G to distill the robustness to the target model. To analyze the impact of the robustness of the teacher model, we compare various teacher models with different dependencies on the attribute G using CIFAR-10B. We consider four teacher models, ordered by robustness to the attribute G: CP ($\boldsymbol{\theta}^T_{\text{CP}}$), Scratch ($\boldsymbol{\theta}^T_{\text{Scratch}}$), and CKD model with a Scratch teacher ($\boldsymbol{\theta}^T_{\text{CKD}}$), and a de-biased model trained on CIFAR-10B balanced for G, *i.e.*, $\alpha = 0.5$, ($\boldsymbol{\theta}^T_{\text{De-biased}}$). Using these teacher models, we report DEO, CD, and RFP of CKDs on the CIFAR-10B dataset in Table 4. We observe that the degree of robustness to the attribute G of the teacher model (*i.e.*, RFP^T) highly correlates to DEO. It is because as the teacher model becomes more robust to G, RFP of the target model gets lower, finally leading to a lower DEO while maintaining fair CD. Namely, these results support our theoretical result again.

Table 4: **Impact of robustness to G of the teacher model.** $\boldsymbol{\theta}^T_{\text{CP}}$, $\boldsymbol{\theta}^T_{\text{CKD}}$, and $\boldsymbol{\theta}^T_{\text{Scratch}}$ are CP, CKD, and Scratch teacher model. $\boldsymbol{\theta}^T_{\text{De-biased}}$ is a Scratch model trained on a perfectly de-biased training dataset ($\alpha = 0.5$). RFP^T denotes how a teacher is biased towards G. CD, DEO, RFP are metrics for evaluating CF, GF, and bias towards G, respectively. Results are measured on CIFAR-10B with $\alpha = 0.8$

Method	RFP^T ↓	Acc ↑	CD ↓	DEO ↓	RFP ↓
CKD w/ $\boldsymbol{\theta}^T_{\text{CP}}$	41.46	76.15	3.59	12.65	18.08
CKD w/ $\boldsymbol{\theta}^T_{\text{Scratch}}$	21.38	78.49	2.85	7.30	11.66
CKD w/ $\boldsymbol{\theta}^T_{\text{CKD}}$	**11.66**	78.39	**2.33**	**4.89**	**5.30**
CKD w/ $\boldsymbol{\theta}^T_{\text{De-biased}}$	**10.81**	77.17	**2.79**	**4.01**	**4.67**

Table 5: **Evaluation of GF and CF of fair-training for image classification**. The details are the same as Table 2. "Scratch" denotes a model trained without considering the notion of fairness through the vanilla cross-entropy loss. "+aug" denotes counterfactual (CTF) image augmentation described in Section 3. If a model performs worse than the Scrath model on CD/DEO, we highlight the numbers in **red**. The best performance is highlighted in **orange**, and the second-best performance is highlighted in **grey**.

Method	CIFAR-10B (α=0.8)			CelebA (and CelebA-CF)			LFW (and LFW-CF)		
	Acc ↑	CD ↓	DEO ↓	Acc ↑	CD ↓	DEO ↓	Acc ↑	CD ↓	DEO ↓
Scratch	78.01	17.90	27.46	95.53	10.26	47.10	90.85	18.06	7.66
SS [14]	74.77	16.42	25.73	95.44	9.13	42.95	90.43	**18.19**	6.75
RW [17]	76.53	12.15	18.94	95.16	5.50	24.21	90.87	**18.68**	6.92
COV [43]	79.03	13.90	24.05	94.42	7.72	34.04	90.85	16.43	6.99
MFD [16]	76.84	12.24	15.39	94.37	4.61	19.00	90.47	16.07	2.15
LBC [15]	76.16	15.01	17.12	94.92	6.24	22.61	90.71	15.76	3.56
SS+aug	73.45	9.95	15.21	95.17	5.24	40.80	89.96	15.23	6.82
RW+aug	76.15	12.93	20.94	95.13	5.34	24.63	90.76	**18.63**	6.71
COV+aug	76.52	8.17	15.04	94.08	8.11	29.03	90.47	13.65	6.78
MFD+aug	77.10	11.16	14.79	93.78	3.87	14.36	89.90	**19.36**	2.47
LBC+aug	75.82	9.01	15.29	94.39	9.32	36.08	88.66	12.41	2.79
CP [36]	75.26	**2.05**	**33.23**	94.10	**2.53**	**51.01**	89.77	9.20	**8.74**
SS+CP	76.54	3.14	**9.08**	94.54	**2.40**	37.97	88.7	**6.13**	4.26
RW+CP	75.68	8.83	13.92	95.19	4.67	25.56	90.87	15.24	6.16
COV+CP	77.74	4.30	19.42	94.29	5.36	**51.63**	91.23	11.91	6.52
MFD +CP	76.67	10.01	13.17	93.81	3.47	23.31	89.39	15.15	**1.90**
LBC+CP	76.88	3.02	12.45	95.12	4.72	22.78	89.92	8.33	3.02
CKD ($\lambda = 0$)	76.32	8.59	11.23	94.12	4.31	**14.11**	90.76	12.42	2.64
CKD	78.49	**2.85**	**7.30**	93.08	4.44	**13.23**	89.26	**7.94**	**1.88**

6 Full comparisons of fair-training methods on image classification

Finally, we evaluate the existing fair-training methods focusing on group fairness (GF) and counterfactual fairness (CF) on CelebA and LFW, together with CIFAR-10B for image classification tasks. We emphasize that only CelebA-CF and LFW-CF have counterfactual images of the real-world images; hence, we measure a CF metric, *i.e.*, Counterfactual Disparity (CD) (Equation (1)), using our datasets. Along with the CF-aware methods, such as CP [36] and CKD, we report the GF-aware methods including SS [14], RW [17], COV [43], MFD [16], and LBC [15]. In addition, we report the naive combinations of GF-aware and CF-aware methods, *e.g.*, training GF-aware method with the augmented training dataset with counterfactual images generated by IP2P [2] (denoted as "+aug") and combinations of the GF-aware methods and the CP regularization (Equation (4)) (denoted as "+CP"). The hyperparameters for all methods besides the GF-aware methods are selected using the same protocol in Section 3, and ones for the GF-aware methods are chosen based on DEO using the same lower bound of the accuracy. Implementation details are provided in Appendix D.3.

Table 5 shows the holistic evaluation of CF and GF for all the methods mentioned above on the three image classification tasks. The table shows four important observations. First, although CP (a CF-aware method) mostly performs the best on CD, it even shows worse DEO than Scratch. We theoretically and empirically discussed the reason in Section 4 and 5. Second, the GF-aware methods are effective in improving DEO but have a minimal impact on CD. This suggests that the faithfulness assumption for SCM may not hold, which will be discussed in more details in Appendix F. Third, the naive combinations of GF-aware and CF-aware methods exhibit much better CD than using the GF-aware methods alone. Additionally, DEOs achieved by the naively combined methods tend to be improved since their training datasets are balanced over the sensitive attributes by incorporating generated samples into the original training datasets. Lastly, we found that CKD shows the best DEO for every evaluation dataset. It shows that if we can train a CF-aware model by reducing the dependency on G, we can achieve both CF and GF even on the image classification task. We additionally conduct an ablation study on CKD by removing CP (*i.e.*, CKD ($\lambda = 0$)) from Equation (6). We observe that this ablated version achieves suboptimal performances than CKD.

This suggests that adding the CP regularization term to the CKD objective function can be helpful to improve both CD and DEO.

7 Concluding remarks

This paper offers carefully crafted benchmark datasets for evaluating the counterfactual fairness (CF) of image classification methods. Since obtaining true counterfactual images is impossible in practice, we employ a high-quality image editing technique to generate counterfactual images of the given images. We construct two facial image benchmarks, CelebA-CF and LFW-CF, by carefully filtering out and verifying the generated counterfactual images by human annotators. Our datasets relax the constraints of the impossibility of evaluating CF in image classification. Using our datasets, we also provide theoretical and empirical results showing that CF may not imply GF, contradictory to the studies conducted on tabular datasets. We elucidate this phenomenon by the presence of the third-party attribute highly correlated with, but not caused by, the sensitive attribute. From this finding, we propose a simple baseline method, CKD, to achieve CF and GF simultaneously. Our extensive experimental results on both GF and CF metrics show that when reducing the reliance on the attribute (*e.g.*, by using CKD), improving the CF metric leads to a significant improvement in the GF metric. By providing our benchmarks and various analyses, we believe that our findings bridge CF and GF in image classification, contributing to the development of fair and robust image recognition systems.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant [No.2021R1A2C2007884] and by Institute of Information & communications Technology Planning & Evaluation (IITP) grants [RS-2021-II211343, RS-2021-II212068, RS-2022-II220113, RS-2022-II220959] funded by the Korean government (MSIT). It was also supported by AOARD Grant No. FA2386-23-1-4079, SNU-Naver Hyperscale AI Center, and Hyundai Motor Chung Mong-Koo Foundation.

References

[1] J. R. Anthis and V. Veitch. Causal context connects counterfactual fairness to robust prediction and group fairness. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023. 1, 2, 5, 6, 21

[2] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 18392–18402, 2023. 2, 9, 13, 17, 27

[3] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conf. Fairness, Accountability and Transparency (FAccT)*, pages 77–91. PMLR, 2018. 1

[4] S. Dash, V. N. Balasubramanian, and A. Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *IEEE/CVF Winter Conf. App. Comput. Vis. (WACV)*, pages 915–924, 2022. 2, 5

[5] M. D'Incà, C. Tzelepis, I. Patras, and N. Sebe. Improving fairness using vision-language driven image augmentation. In *IEEE/CVF Winter Conf. App. Comput. Vis. (WACV)*, pages 4695–4704, 2024. 2

[6] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *Int. Conf. Mach. Learn. (ICML)*, pages 2803–2813. PMLR, 2020. 4

[7] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. In *AAAI/ACM Conf. AI, Ethics, and Society (AIES)*, pages 219–226, 2019. 2, 4, 5, 18, 23

[8] M. Goldblum, L. Fowl, S. Feizi, and T. Goldstein. Adversarially robust distillation. In *Proc. of the AAAI Conf. Artificial Intelligence (AAAI)*, volume 34, pages 3996–4003, 2020. 21

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 27, 2014. 2

[10] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, volume 29, 2016. 5

[11] E. Harlan and O. Schnuck. Objective or biased – the questionable use of artificial intelligence in job applications. *bayerischer rundfunk*, 2021. URL https://interaktiv.br.de/ki-bewerbung/en/. 1

[12] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 7, 21, 24

[13] G. B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Int. Conf. Comput. Vis. (ICCV)*, 2007. 3, 13, 17, 18, 26

[14] B. Y. Idrissi, M. Arjovsky, M. Pezeshki, and D. Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conf. Causal Learning and Reasoning (CLeaR)*, pages 336–351. PMLR, 2022. 9, 19, 23

[15] H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 702–712. PMLR, 2020. 9, 19, 23

[16] S. Jung, D. Lee, T. Park, and T. Moon. Fair feature distillation for visual recognition. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 12115–12124, 2021. 7, 9, 19, 23

[17] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems (KAIS)*, 33(1):1–33, 2012. 9, 19, 23

[18] R. Kennaway. When causation does not imply correlation: Robust violations of the faithfulness axiom. In *The Interdisciplinary Handbook of Perceptual Control Theory*, pages 49–72. Elsevier, 2020. 21

[19] H. Kim, S. Shin, J. Jang, K. Song, W. Joo, W. Kang, and I.-C. Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proc. of the AAAI Conf. Artificial Intelligence (AAAI)*, volume 35, pages 8128–8136, 2021. 2

[20] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 31, 2018. 5

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[22] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017. 2, 5

[23] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2017. 1

[24] H. Liang, P. Perona, and G. Balakrishnan. Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. In *Int. Conf. Comput. Vis. (ICCV)*, pages 4977–4987, 2023. 2

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Int. Conf. Comput. Vis. (ICCV)*, pages 3730–3738, 2015. 3, 13, 17, 18, 26

[26] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 30, 2017. 2

[27] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023. 27

[28] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 8, 19

[29] L. S. Nguyen and D. Gatica-Perez. Hirability in the wild: Analysis of online conversational video resumes. *IEEE Trans. Multimedia*, 18(7):1422–1437, 2016. 1

[30] J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000. 5

[31] M. Peychev, A. Ruoss, M. Balunović, M. Baader, and M. Vechev. Latent space smoothing for individually fair representations. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 535–554. Springer, 2022. 4, 5, 18, 19, 23

[32] S. R. Pfohl, T. Duan, D. Y. Ding, and N. H. Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference (MLHC)*, pages 325–358. PMLR, 2019. 2

[33] M. Pinto, A. V. Carreiro, P. Madeira, A. Lopez, and H. Gamboa. The matrix reloaded: Towards counterfactual group fairness in machine learning. *Journal of Data-centric Machine Learning Research (DMLR)*, 2024. 4, 22

[34] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 9301–9310, 2021. 2

[35] L. Rosenblatt and R. T. Witter. Counterfactual fairness is basically demographic parity. In *Proc. of the AAAI Conf. Artificial Intelligence (AAAI)*, volume 37, pages 14461–14469, 2023. 1, 2, 5, 21

[36] C. Russell, M. J. Kusner, J. Loftus, and R. Silva. When worlds collide: integrating different counterfactual assumptions in fairness. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 30, 2017. 2, 4, 5, 6, 9, 18, 19, 21, 23, 24

[37] L. Scimeca, S. J. Oh, S. Chun, M. Poli, and S. Yun. Which shortcut cues will dnns choose? a study from the parameter-space perspective. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 7

[38] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 7

[39] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Int. Conf. Comput. Vis. (ICCV)*, pages 692–702, 2019. 1

[40] H. Wu, G. Bezold, M. Günther, T. Boult, M. C. King, and K. W. Bowyer. Consistency and accuracy of celeba attribute values. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 3257–3265, 2023. 3

[41] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018. 2

[42] M. Yurochkin and Y. Sun. Sensei: Sensitive set invariance for enforcing individual fairness. *arXiv preprint arXiv:2006.14168*, 2020. 4, 5, 18, 19, 23

[43] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, pages 962–970. PMLR, 2017. 1, 9, 19, 23

[44] F. Zhang, K. Kuang, L. Chen, Y. Liu, C. Wu, and J. Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *Int. Conf. Learn. Represent. (ICLR)*, 2022. 2

[45] B. Zi, S. Zhao, X. Ma, and Y.-G. Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Int. Conf. Comput. Vis. (ICCV)*, pages 16443–16452, 2021. 7

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section H.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Our claims are reflected accurately.
   (b) Did you describe the limitations of your work? [Yes] See Appendix B
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Appendix D
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix G.1
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix D.2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited Liu et al. [25], Huang et al. [13] and Brooks et al. [2].
   (b) Did you mention the license of the assets? [Yes] See Appendix C.3
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Appendix H
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We did not create new data but edited existing image benchmarks, so this issue does not apply.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We did not create new data but edited existing image benchmarks, so this issue does not apply.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Figure C.1, C.2, and C.3.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] See Appendix C.2

A Proof of Theorem 4.1

We start from LHS in equation 3:

$$
\begin{aligned}
&\left|P(\widehat{Y}=y'|A=0,Y=y)-P(\widehat{Y}=y'|A=1,Y=y)\right| \\
=&\Bigg|\sum_{X_A,X_Y,X_G} P(\widehat{Y}=y'|X_A,X_Y,X_G,A=0,Y=y)P(X_A,X_Y,X_G|A=0,Y=y) \\
&\qquad -P(\widehat{Y}=y'|X_A,X_Y,X_G,A=1,Y=y)P(X_A,X_Y,X_G|A=1,Y=y)\Bigg| \qquad \text{(A.1)} \\
=&\Bigg|\sum_{X_A,X_Y,X_G} P(\widehat{Y}=y'|X_Y,X_G)P(X_A,X_Y,X_G|A=0,Y=y) \\
&\qquad -P(\widehat{Y}=y'|X_Y,X_G)P(X_A,X_Y,X_G|A=1,Y=y)\Bigg| \qquad \text{(A.2)} \\
=&\Bigg|\sum_{X_Y,X_G} P(\widehat{Y}=y'|X_Y,X_G)\big(P(X_Y|X_G,A=0,Y=y)P(X_G|A=0,Y=y) \\
&\qquad -P(X_Y|X_G,A=1,Y=y)P(X_G|A=1,Y=y)\big)\Bigg| \qquad \text{(A.3)} \\
=&\Bigg|\sum_{X_Y,X_G} P(\widehat{Y}=y'|X_Y,X_G)\Big(P(X_Y|Y=y)P(X_G|A=0,Y=y)-P(X_Y|Y=y)P(X_G|A=1,Y=y)\Big)\Bigg| \qquad \text{(A.4)} \\
=&\Bigg|\sum_{X_Y} P(X_Y|Y=y)\Big(\sum_{X_G} P(\widehat{Y}=y'|X_Y,X_G)P(X_G|A=0,Y=y) \qquad \text{(A.5)} \\
&\qquad -\sum_{X'_G} P(\widehat{Y}=y'|X_Y,X'_G)P(X'_G|A=1,Y=y)\Big)\Bigg|. \qquad \text{(A.6)}
\end{aligned}
$$

Note the first and third equalities are driven by Bayes' theorem, the second one is from the independence between $\widehat{Y}$ and X_G, A conditioned on X_Y, X_G based on the Markov properties of SCM, and the fourth one is due to the independence between X_Y and X_G, A conditioned on Y. We denote a coupling between the two distributions $P(X_G|A=0,Y)$ and $P(X'_G|A=1,Y)$ as $\Pi(X_G, X'_G)$, then we have:

$$
\begin{aligned}
&\left|P(\widehat{Y}=y'|A=0,Y=y)-P(\widehat{Y}=y'|A=1,Y=y)\right| \\
=&\Bigg|\sum_{X_Y} P(X_Y|Y=y)\Big(\sum_{X_G,X'_G}\Pi(X_G,X'_G)\big(P(\widehat{Y}=y'|X_Y,X_G)-P(\widehat{Y}=y'|X_Y,X'_G)\big)\Big)\Bigg|. \qquad \text{(A.7)} \\
\leq&\sum_{X_Y} P(X_Y|Y=y)\Big(\sum_{X_G,X'_G}\Pi(X_G,X'_G)\Big|P(\widehat{Y}=y'|X_Y,X_G)-P(\widehat{Y}=y'|X_Y,X'_G)\Big|\Big) \qquad \text{(A.8)} \\
=&\sum_{X_Y} P(X_Y|Y=y)\sum_{X_G,X'_G}\Pi(X_G,X'_G)d_{\boldsymbol{\theta},X_Y}(X_G,X'_G) \qquad \text{(A.9)}
\end{aligned}
$$

where the sample distance is denoted as $d_{\boldsymbol{\theta},X_Y}(X_G,X'_G)=\left|P(\widehat{Y}=y'|X_Y,X_G)-P(\widehat{Y}=y'|X_Y,X'_G)\right|$. The inequality in Equation A.8 is driven by Jensen's inequality.

B Limitations and societal impacts

While our datasets and analyses reveal the relationship between CF and GF in image classification, we clarify our study's limitations. First of all, our study uses sex as a sensitive attribute based on visually perceived biological traits. However, as mentioned in Section 2, this simplification does not capture the full spectrum of sexual traits, which is more complex and nuanced. Therefore, we emphasize again that practitioners should use our data with these considerations in mind; they should not utilize our datasets for gender categorization but rather for investigating the unfairness in terms of CF and GF and enhancing fairness in AI systems. Second, our data generation process relies on IP2P to create CTF samples. We tried to mitigate the potential bias problem during the data generation process through the sophisticated human filtering process, but our data samples could be

Figure A.1: **LFW-CF examples**. The counterfactual (CTF) images regarding the "sex attribute" are shown. The top row shows the original image, while the bottom row displays the CTF image generated by IP2P.

Figure A.2: **CF examples with other sensitive attributes**. Original and CTF samples are shown when age or skin color is considered as the sensitive attribute.

affected by the unintended bias of IP2P. Third, while we assume the structural causal model (SCM) for images as Figure 2, specifying an SCM in the real world is often infeasible. This difficulty also makes it challenging to apply some of our experiments, such as analyzing the attribute G or using a robust teacher model to G. However,

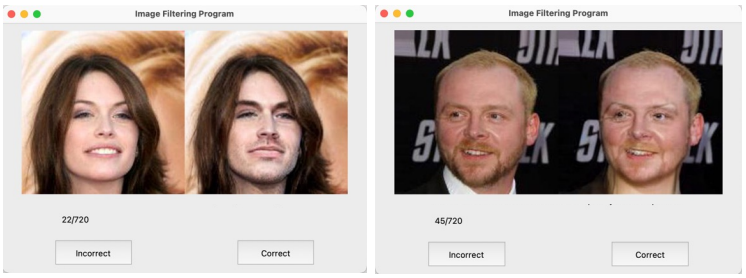

Figure C.1: User Interface shown to five annotators for image filtering.

we addressed these challenges to some extent by proposing a systematic method to investigate G (Appendix E.2) and analyzing the robustness of vanilla teacher models (Section 5.3).

Despite the limitations, we believe that our work makes significant contributions through our curated image datasets and extensive analyses to the evolving field that addresses relationships among different fairness notions.

C Datasets: CelebA-CF and LFW-CF

We provide access to our newly created dataset, *i.e.*, CelebA-CF and LFW-CF, through the following link: CelebA-CF (https://figshare.com/s/62b6f7f69d0eab9c3c71) , LFW-CF (https://figshare.com/s/39f2daac58148e10e5fe)

C.1 Hyperparameters for IP2P image editing

We set the resolution of generated images to 256×256 and the denoising step to 50. Furthermore, we set the Image-CFG weight to 1.8 and the Text-CFG weight to 7.5. These two hyperparameters are the guidance scales to control how the generated images closely resemble the input image or are intensely edited. To alter the sensitive attribute of facial images, we use the prompts of "turn the woman into a man" for female images and "turn the man into a woman" for male images.

C.2 Human annotations

After generating CTF images using IP2P, we filtered them through five annotators to construct high-quality CTF samples, namely CelebA-CF and LFW-CF. Before evaluating the created CTF samples, the guidelines are given to the annotators, as presented in Figure C.2. To establish the guidelines, we extracted 20 facial attributes of secondary sex characteristics using Chat-GPT and then, with guidance from experts specialized in fairness, selected 9 key facial attributes (facial hair, Adam's apple, skin texture, jawline, chin shape, brow ridge, cheekbone prominence, lip fullness, and hairline). The guidelines instruct human annotators to filter out counterfactual samples based on these attributes (including considerations for the presence of makeup). Subsequently, given the original image and generated CTF image pair, annotators assess whether the generated image is correctly created based on the instructions. Figure C.1 shows the interfaces of the annotation task for image filtering.

We further verify the reliability of our datasets with another five human annotators, different from those who participated in the previous filtering process, and report the result in Table 1. For more objective annotation, we show 8 example images to the annotators before the labeling task, which are randomly sampled the same number of times for each attribute value from test image datasets. Then, the annotators label CelebA-CF and LFW-CF for 4 attributes, *i.e.*, "sex", "blond hair", "gray hair", and "smiling" in order. Specifically, the annotators evaluate whether the sensitive attribute was correctly altered and the non-sensitive attributes were maintained for a generated image. Similar to the image filtering task, we provide the annotators with the set of 10 sex-related facial attributes for objective and accurate labeling. Figure C.3 shows the interfaces of the annotation task for this reliability check. We provide the attribute values originally annotated for the original image datasets on the screen together for annotators to refer to as a guide for their annotating tasks. Although we focus on visually perceived sex traits, we use the terms Male and Female for convenience in the annotation interface.

We note that the wage paid to each participant is 18 USD per hour, resulting in a total expenditure of 360 USD on participant compensation.

Evaluation Guidelines for Generated Images

- **Instructions**:
  - You will receive a pair of images: an original real image and a generated CTF image.
  - Your task is to evaluate the generated CTF image and label it as either "correct" or "incorrect".
- **Key Rule**:
  - Label the generated CTF image as "correct" if it maintains the target attribute while showcasing a change in "biological secondary sex characteristics + make-up" compared to the original image.
  - Otherwise, label it as "incorrect".
- **Cases for Incorrectness**:
  1. Presence of distorted or unrealistic facial features in the generated image.
  2. Significant alterations beyond "biological secondary sex characteristics + make-up" (e.g., hair style, skin color, etc.).
  3. Absence of change in "biological secondary sex characteristics + make-up" change in the generated image.

A change in **biological secondary sex characteristics + make-up** refers to a visible difference in one or more of the following features:

- **biological secondary sex characteristics:**
  - Facial hair
  - Adam's apple
  - Skin texture
  - Jawline
  - Chin shape
  - Brow ridge
  - Cheekbone prominence
  - Lip fullness
  - Hairline
- **Makeup**

Figure C.2: The guideline instructions that were given to the five annotators for the image filtering.

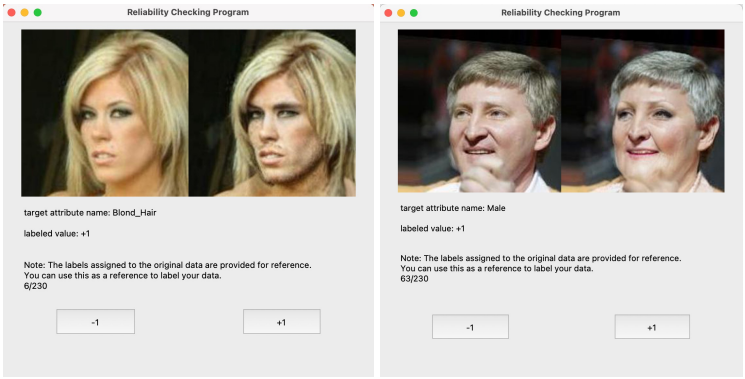

Figure C.3: User Interface shown to five annotators to evaluate the reliability of our created datasets.

C.3 License information of assets employed in this study

- CelebA [25] was made available for academic research purposes without a formal license. The dataset can be downloaded at https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.
- LFW [13] is publicly available for research purposes. There is also no formal license and further information is reported at https://vis-www.cs.umass.edu/lfw/.
- InstructPix2Pix (IP2P) [2] is licensed under the MIT license and is available at https://github.com/timothybrooks/instruct-pix2pix.
- IP2P further builds upon stable-diffusion-v1-5 that is released under CreativeML- Open-RAIL-M License.

C.4 Further information of the new dataset.

CelebA-CF and LFW-CF are based on real facial image datasets, such as CelebA and LFW, which include 40 attribute annotations. We note that because the original datasets have an imbalance between some pairs of two attributes, our datasets also possess a different skewness between the sensitive attribute and other attributes. For example, in CelebA and CelebA-CF, most images with the blond hair attribute are female, and most males do not have blond hair. We present the skewness values between the sensitive attribute and other non-sensitive attributes in CelebA-CF in Table C.1. Additionally, we displayed the group-specific failure rate identified through the filtering process in Table C.2.

Table C.1: **The skewness between the sensitive attribute and other non-sensitive attributes in CelebA-CF.**

Attribute	Hair Length	Bangs	Wearing Hat	Brown Hair	Pale Skin	Big Lips	Mouth Slightly Open	Smiling	Wavy Hair
Skewness	0.798	0.907	0.974	0.820	0.980	0.815	0.265	0.335	0.833

Table C.2: **The failure rate identified through the filtering process.** This table shows the proportion of images filtered out, calculated separately for each sensitive attribute (*e.g.*, female, male), in constructing the CelebA-CF and LFW-CF datasets. The sensitive attributes in the table represent the labels of the original images.

	CelebA-CF	LFW-CF
Female	0.57	0.70
Male	0.80	0.69

D Implementation details

D.1 Details on training datasets

For CelebA, we utilize the official train-validation-test split [25]. For LFW, we also use the official train-test split [13] and then divide the training data into a training and a validation set, with a ratio of 80:20.

CIFAR-10B is a modified dataset from CIFAR-10, as described in Section 5.2. We modify CIFAR-10 into a binary classification task by dividing the original 10 classes into two classes (classes 0-4 and 5-9). To introduce a fairness issue, we set the sensitive attribute A as the presence of Gaussian noise and skew the dataset by randomly injecting the noise into 20% and 80% of the data in class 0 and class 1, respectively. Additionally, we introduce Contrast noise for the attribute G. Using the skew-ratio α, we create a statistical correlation between A and G by adding the noise into $100 \times \alpha\%$ of the data samples with $A = 1$ and $100 \times (1 - \alpha)\%$ of the data samples with $A = 0$. Unless otherwise noted, we set α to 0.8. We partition the dataset into train-validation-test sets with a ratio of 64:16:20, respectively, maintaining consistent values for the two skewness ratios (*i.e.*, skewness between A and Y, G and A) across all sets during our experiments.

D.2 Compute Infrastructure and optimization

Our all experiments including the dataset construction and performance comparison of existing methods and CKD were conducted using AMD Ryzen Threadripper PRO 3975WX CPUs and NVIDIA RTX A5000 GPUs. Our dataset generation was parallelized using 8 GPUs and took 2 days to complete. Training time for models used in the experiments for performance comparison ranges between 12 to 24 hours depending on the dataset and method used.

For CIFAR-10B, we use ResNet56 models with the Adam optimizer for 50 epochs. We set the mini-batch size and learning rate as 128 and 0.001, respectively. Because the skewness between G and A in CIFAR-10B test datasets varies, we compute the balanced CD over both the target class and the sensitive attribute for the consistent metric. For CelebA and LFW, we train ResNet18 models with the AdamW optimizer. We use the epoch size of 70 and 50 for each dataset, and set the mini-batch size, learning rate, and weight decay as 128, 0.001, and 1e-4, respectively. We use identical hyperparameters regarding the optimization for all methods. All results are averaged over results from four different random seeds.

D.3 Implementation details of baselines and CKD

CF-aware methods. Scratch (+Aug) [7] minimizes the empirical cross-entropy loss computed using both original and counterfactual images. CP [36] has a regularization term that promotes the image pairs to be the same prediction. We use logits of a neural network model as representation vectors for the CP regularization term. Since Scratch (+aug) and CP utilize CTF samples, we generated these samples using IP2P with the same prompt in Appendix C.1 using the image-CFG of 7.5 and the Text-CFG of 2.0. We note that we do not apply any filtering process for their generated training datasets. SenSeI [42] uses two metrics for training: one for a pre-defined fair regularizer distance metric and the other obtained by fair metric learning. We use the same metrics as presented in their code. By generating the worst-case samples based on these metrics, we apply a fair regularization term to promote their predictions to be the same, as originally implemented. LASSI [31] minimizes an objective function which is composed of the classification loss, the reconstruction loss, and the adversarial loss to learn individually fair representation. We use the official code of LASSI as it is.

GF-aware methods. LBC [15] necessitates multiple full-training iterations, alternately re-weighting each group based on the given group fairness metric and re-training. Due to its high computation budget for iterative full-training, we limit the number of epochs for each training to 5 and repeat this process 14 times. COV [43] utilizes a fairness constraint based on the covariance between the group label and the signed distance of feature vectors from the decision boundary of a classifier. We minimize the constraint-regularized objective function through gradient descent optimization, instead of directly solving its optimization problem. MFD [16] employs an additional fairness-promoting regularization term based on Maximum Mean Discrepancy (MMD). For the MMD distance of the regularization term, we use the Gaussian RBF kernel with the variance parameter set as the mean of squared distance between all data points. We implemented SS [14] and RW [17] identically to the original algorithm.

CF- and GF-aware methods. The combinations of GF method and the augmentation were implemented so that GF methods train a model on their own objective function using training datasets augmented by generated CTF samples. The combinations of GF methods and CP optimize the objective functions of GF methods combined by the CP regularization. The CKD regularization term (5) builds upon representation vectors $f(\boldsymbol{\theta}, x)$. For this vector, we use the logits of a neural network model on LFW and CIFAR-10B. For CelebA, we utilize feature vectors from the penultimate layer of models as a representation vector since its training dataset is relatively much larger and more complex than others, leading to more fine-grained feature vectors.

Our code is available at https://github.com/sumin-yu/CKD.git.

D.4 Hyperparameter search

The range of hyperparameter search used for Table 2 and Table 5 are shown in Table D.1. We utilize grid search to select hyperparameter values within a certain range. Note CKD and the combinations of GF-aware methods and CP have additional parameters λ for CP loss. We use the same range as CP for λ. We also note that for LASSI, we only search the hyperparameter for an adversarial loss while maintaining other parameters as the same as used in their experiments on CelebA.

Table D.1: Hyperparameters and search ranges for each method.

Method	Hyperparameter	Search range
CP [36]	CP strength λ	$[10^{-2}, 10^{2}]$
SenSeI [42]	Fair regularization strength ρ	$[10^{-2}, 10^{2}]$
LASSI [31]	Adversarial loss weight λ_2	$[10^{-3}, 10^{-1}]$
COV [43]	Covariance strength λ	$[10^{-2}, 10^{2}]$
MFD [16]	MMD strength λ	$[10^{-1}, 10^{6}]$
LBC [15]	LR for re-weights η	$[10^{-1}, 10^{3}]$
CKD	CKD strength μ	$[10^{-1}, 10^{3}]$

E Rate of Flipped Predictions (RFP)

E.1 RFP measurement on CelebA

To measure RFP on CelebA, we assume that the hair length of facial images is G, *i.e.*, it is correlated with, but not caused by, the sensitive attribute. Then, we use Stochastic Differential Editing (SDEdit) [28], an image editing method based on a diffusion model, to modify the hair length in each image. SDEdit selectively edits specific regions of a given image based on the colored stroke. Namely, SDEdit depicts the image region indicated by the stroke with the given color in the most plausible manner. By doing so, SDEdit generates realistic and faithful edited images, while preventing changes in the region not indicated by the strokes. To utilize SDEdit, we randomly select 40 samples for each group of the same target label and sensitive attribute from original samples of CelebA-CTF pairs and then we manually apply strokes on the hair of facial images for a total of 160 samples. Specifically, To extend the hair length, we applied strokes with the hair color to the areas where the hair should grow. Conversely, to shorten the hair length, we applied strokes with the background color to the areas where the hair should be removed. After this process, we utilize the official PyTorch implementation of Meng et al. [28] to edit images with the applied strokes. Figure E.1 shows some examples of images edited by SDEdit.

Table E.1: The skewness between the sensitive attribute and other attributes, as well as the accuracy for each attribute after re-training a linear classifier on the top of the CP-trained model.

	Hair Length	Bangs	Wearing Hat	Pale Skin	Mouth Slightly Open
Acc (%)	64.93	68.07	72.63	50.79	56.79
Skewness	0.8	0.91	0.97	0.98	0.27

Figure E.1: **Examples of CelebA measuring RFP with respect to "hair length".** If the person in the original image (first row) had long hair, we create a modified image (second row) with shorter hair and vice versa.

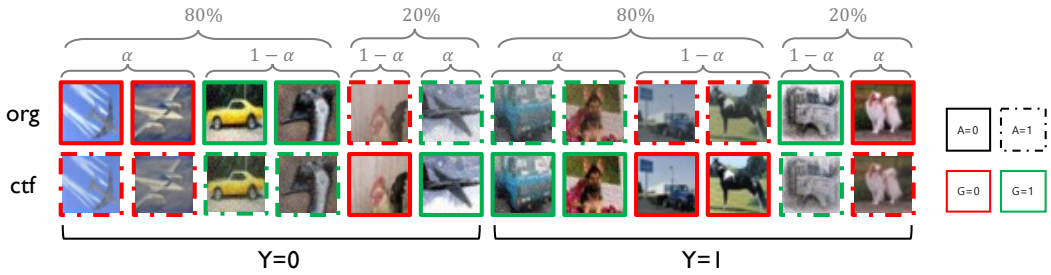

Figure E.2: **Illustration of CIFAR-10B.** The sensitive attribute A is characterized by the line type ($A = 0$ for a solid line, $A = 1$ for a dashed line), while another attribute G is denoted by the color of the line ($G = 0$ for a red line, $G = 1$ for a green line).

E.2 Discussion about selecting G on CelebA

As mentioned in Section 5.2, we intuitively chose "hair length" as G on CelebA since G is highly correlated with, but not caused by, the sensitive attribute A. However, we introduce a more generalized approach for choosing G by leveraging a CP-trained model that exhibits low CD but high DEO. Specifically, using a pre-defined set of attributes, we first train a linear classifier on the top of the feature extractor from the CP-trained model for each attribute. In cases where annotations are not available, CLIP-based pseudo labels can be utilized. Based on the accuracy of each linear classifier, we can then identify which attributes the CP-trained model learns more, indicating potential heavy reliance on these attributes. Finally, we can select G attributes based on two criteria: (1) high accuracy of a linear classifier and (2) high correlation (not causation) with the sensitive attribute.

To validate this approach, we conducted an experiment on CelebA dataset using a subset of 40 pre-annotated attributes and the "hair length" ("hair length" labels are predicted by CLIP as it is not originally labeled in CelebA). Table E.1 displays the accuracy for each attribute after training a linear classifier on the top of the CP-trained model, alongside the skewness between the sensitive attribute and other attributes. Attributes like "Pale Skin" show a high correlation with the sensitive attribute but low accuracies, suggesting CP might not rely on them (not satisfying the second condition). "Mouth Sightly Open" exhibits low correlation and accuracy, thus not being considered as G (failing the first condition). In contrast, attributes such as "bangs", "wearing hat", and "hair length" exhibit both high correlation values and accuracies, indicating that they are promising candidates for the attribute G.

F Further discussion about the faithfulness assumption

The faithful assumption states that if some two variables are statistically independent, they are d-separated, *i.e.*, there is no connected path between them. Thus, under this assumption, GF-aware methods, which enforce independence between the sensitive attribute and the target label, can achieve CF simultaneously when they successfully achieve GF. The previous works [1, 35] provide the same argument that GF implies CF under the faithfulness assumption. However, some other previous works [18] have demonstrated that while the faithfulness assumption is crucial for causal inference literature, it may not always hold true, especially in complex real-world scenarios. Moreover, the result for GF-aware methods in Table 5 reveals that the GF-aware methods have a minimal impact on improving CD, implying that the faithfulness assumption does not hold in CelebA and LFW datasets.

G Additional results

G.1 Result tables with standard deviation

We report the standard deviation values of the performance comparison results in Table G.1 for CIFAR-10B, CelebA, and LFW respectively. The standard deviation values are calculated over four different seeds.

G.2 Impact of the robustness to G of the teacher model on CKD with CelebA and LFW

We analyze the effectiveness of the teacher model on real image datasets. We report the performance of CKD using variants of teacher models that are more or less robust to G on CelebA and LFW. We consider three teacher models, ordered by robustness to G: CP ($\boldsymbol{\theta}^T_{\text{CP}}$), Scratch ($\boldsymbol{\theta}^T_{\text{Scratch}}$), and CKD model with a Scratch teacher ($\boldsymbol{\theta}^T_{\text{CKD}}$). Table G.2 displays ACC, DEO, CD (on CelebA and LFW), and RFP (on CelebA) depending on the teacher models. Since we generate a dataset for RFP measurements on CelebA with "hair length" as G, we report RFP values only for CelebA. Through the result, we observe that compared to vanilla-trained teachers $\boldsymbol{\theta}^T_{\text{Scratch}}$, using more robust teachers (*e.g.*, $\boldsymbol{\theta}^T_{\text{CKD}}$) achieves slightly better or competitive DEO and CD, while employing less robust teachers (*e.g.*, $\boldsymbol{\theta}^T_{\text{CP}}$) significantly degrades DEO, which are consistent with the results in Section 5.3 on CIFAR-10B.

G.3 Analaysis on CKD

Ablation study. To study the effectiveness of our CKD regularization term, we additionally consider a method that is a naive combination of a typical KD method proposed by Hinton et al. [12] and CP [36] (*i.e.*, HKD+CP). Note that this combination can be considered as a baseline method that considers both CF and GF if it uses a robust teacher model to G, because the method robustifies the model with respect to G while achieving CF. Table G.5 compares the method with our CKD. As we expected, the results show that HKD+CP improves both DEO and CD simultaneously. However, its performance is still suboptimal compared to CKD, showing CKD is more effective than the naive combination of KD and CP.

CKD on feature vectors vs logits. CKD can utilize either feature or logit vectors as the target vectors, i.e., $f(\theta, x)$. For CIFAR-10B and LFW, where we use logits as the target vectors in our experiments, we displayed the performance of CKD using feature vectors as the target vectors in Table G.6. The results demonstrate that CKD using feature vectors exhibits comparable performance to those with logits. Moreover, we can get even better performance on CIFAR-10B using feature vectors, demonstrating that CF and GF can be achieved simultaneously regardless of which type of target vectors is used.

Sensitiveness of μ. μ is the regularization strength for the CKD loss term. Specifically, as μ increases, we expect improvements in both DEO and CD. Table G.7 shows the performance of CKD across different values of μ, aligning with our expectations. Additionally, we note that CKD performance is insensitive to μ.

The implication of using non-curated IP2P. Uncurated CTF datasets are imperfect. Specifically, some samples generated from the original images in our test datasets were filtered out because the images either showed minimal changes or had alterations that affected non-sensitive attributes including the target attributes. Consequently, the more such incomplete samples exist, the more they will negatively impact the performance of our method. To assess how sensitive CKD is to incomplete CTF training samples, we conducted additional experiments on CIFAR-10B by varying the ratio of incomplete CTF samples in the training set. For a given ratio α , we assumed that half of the incomplete samples are nearly unchanged, while the other half are samples where both the target and sensitive attributes are altered. We varied α from 20% to 60% in 10% increments and reported the accuracy, CD, and DEO of CKD in the table below. The results in Table G.3 indicate that CKD significantly improves both CD and DEO compared to Scratch, even for high α s. Although this phenomenon has not been fully explained, we hypothesize that the robustness can be attributed to the distillation process, as empirically demonstrated in [8].

G.4 Additional experimental results on counterfactual samples.

We additionally report the accuracy (acc-CTF) and DEO (DEO-CTF) for CKD and several baseline methods on counterfactual samples in CelebA-CF in Table G.4.

G.5 Additional experimental results for other metrics proposed by [33].

We first note that the Counterfactual Disparity (CD) we used is the same metric as the Switch Rate (SR) proposed by [33]. We computed P2NR (another metric proposed by [33]) on CelebA and obtained values of 0.036, 0.165, and 0.339 for Scratch, CP, and CKD, respectively. These results indicate that CKD achieves low CD with a balanced rate of misclassification across the labels. Additionally, we would like to emphasize that Pinto et al. focused on scenarios where GF does not imply CF in their experiments—highlighting cases where the faithfulness assumption, which can be overly stringent, does not hold (see line 323 in their paper). However, our work primarily explores the converse: whether CF can imply GF depending on the presence of G, independent of the faithfulness assumption.

Table G.1: **Evaluation of GF and CF of fair-training for image classification**. We put the results from Table 2 and Table 5 together with the standard deviation values over four different seeds.

(a) Standard deviations on CIFAR-10B.

Method	CIFAR-10B (α=0.8) Acc ↑	CD ↓	DEO ↓
Scratch	$78.01_{\pm 0.85}$	$17.90_{\pm 2.43}$	$27.46_{\pm 2.65}$
CF-aware training			
Scratch+aug [7]	$75.38_{\pm 0.79}$	$9.33_{\pm 0.88}$	$15.25_{\pm 1.19}$
CP [36]	$75.26_{\pm 1.87}$	$2.05_{\pm 0.15}$	$33.23_{\pm 7.56}$
SenSeI [42]	$77.21_{\pm 0.83}$	$16.32_{\pm 1.23}$	$24.18_{\pm 1.70}$
GF-aware training			
SS [14]	$74.77_{\pm 0.41}$	$16.42_{\pm 1.34}$	$25.73_{\pm 2.12}$
RW [17]	$76.53_{\pm 0.57}$	$12.15_{\pm 1.23}$	$18.94_{\pm 1.71}$
COV [43]	$79.03_{\pm 0.49}$	$13.90_{\pm 0.39}$	$24.05_{\pm 1.09}$
MFD [16]	$76.84_{\pm 0.92}$	$12.24_{\pm 1.51}$	$15.39_{\pm 1.28}$
LBC [15]	$76.16_{\pm 1.36}$	$15.01_{\pm 2.26}$	$17.12_{\pm 4.68}$
both CF and GF-aware training			
SS+aug	$73.45_{\pm 0.46}$	$9.95_{\pm 0.58}$	$15.21_{\pm 2.06}$
RW+aug	$76.15_{\pm 0.52}$	$12.93_{\pm 0.96}$	$20.94_{\pm 2.57}$
COV+aug	$76.52_{\pm 0.56}$	$8.17_{\pm 0.22}$	$15.04_{\pm 0.96}$
MFD+aug	$77.10_{\pm 0.58}$	$11.16_{\pm 1.08}$	$14.79_{\pm 2.15}$
LBC+aug	$75.82_{\pm 0.54}$	$9.01_{\pm 0.67}$	$15.29_{\pm 0.56}$
SS+CP	$76.54_{\pm 1.48}$	$3.14_{\pm 0.26}$	$9.08_{\pm 1.22}$
RW+CP	$75.68_{\pm 0.37}$	$8.83_{\pm 0.66}$	$13.92_{\pm 0.94}$
COV+CP	$77.74_{\pm 0.52}$	$4.30_{\pm 0.33}$	$19.42_{\pm 1.74}$
MFD+CP	$76.67_{\pm 1.00}$	$10.01_{\pm 0.57}$	$13.17_{\pm 0.95}$
LBC+CP	$76.88_{\pm 2.64}$	$3.02_{\pm 1.29}$	$12.45_{\pm 1.93}$
CKD ($\lambda = 0$)	$76.32_{\pm 0.59}$	$8.59_{\pm 1.32}$	$11.23_{\pm 1.04}$
CKD	$78.49_{\pm 0.66}$	$2.85_{\pm 0.20}$	$7.30_{\pm 0.46}$

(b) Standard deviations on CelebA and LFW.

Method	CelebA (and CelebA-CF) Acc ↑	CD ↓	DEO ↓	LFW (and LFW-CF) Acc ↑	CD ↓	DEO ↓
Scratch	$95.53_{\pm 0.06}$	$10.26_{\pm 1.33}$	$47.10_{\pm 5.57}$	$90.85_{\pm 0.27}$	$18.06_{\pm 1.89}$	$7.66_{\pm 0.49}$
CF-aware training						
Scratch+aug [7]	$95.41_{\pm 0.15}$	$4.65_{\pm 0.86}$	$44.71_{\pm 2.57}$	$90.34_{\pm 0.58}$	$12.15_{\pm 1.29}$	$7.86_{\pm 1.70}$
CP [36]	$94.10_{\pm 0.08}$	$2.53_{\pm 1.26}$	$51.01_{\pm 1.71}$	$89.77_{\pm 0.90}$	$9.20_{\pm 0.56}$	$8.74_{\pm 1.43}$
SenSeI [42]	$95.33_{\pm 0.30}$	$8.00_{\pm 0.59}$	$52.32_{\pm 5.26}$	$87.75_{\pm 3.78}$	$16.09_{\pm 3.70}$	$9.23_{\pm 1.38}$
LASSI [31]	$91.07_{\pm 0.27}$	$9.69_{\pm 0.78}$	$31.79_{\pm 3.17}$	-	-	-
GF-aware training						
SS [14]	$95.44_{\pm 0.09}$	$9.13_{\pm 2.73}$	$42.95_{\pm 3.87}$	$90.43_{\pm 0.21}$	$18.19_{\pm 1.27}$	$6.75_{\pm 0.33}$
RW [17]	$95.16_{\pm 0.08}$	$5.50_{\pm 0.46}$	$24.21_{\pm 1.76}$	$90.87_{\pm 0.25}$	$18.68_{\pm 2.91}$	$6.92_{\pm 0.96}$
COV [43]	$94.42_{\pm 0.16}$	$7.72_{\pm 2.89}$	$34.04_{\pm 4.43}$	$90.85_{\pm 0.50}$	$16.43_{\pm 2.12}$	$6.99_{\pm 1.23}$
MFD [16]	$94.37_{\pm 0.77}$	$4.61_{\pm 1.77}$	$19.00_{\pm 6.44}$	$90.47_{\pm 0.10}$	$16.07_{\pm 2.01}$	$2.15_{\pm 0.51}$
LBC [15]	$94.92_{\pm 0.28}$	$6.24_{\pm 0.69}$	$22.61_{\pm 1.79}$	$90.71_{\pm 0.61}$	$15.76_{\pm 0.86}$	$3.56_{\pm 2.02}$
both CF and GF-aware training						
SS+aug	$95.17_{\pm 0.02}$	$5.24_{\pm 1.02}$	$40.80_{\pm 2.86}$	$89.96_{\pm 0.24}$	$15.23_{\pm 2.21}$	$6.82_{\pm 0.88}$
RW+aug	$95.13_{\pm 0.04}$	$5.34_{\pm 0.59}$	$24.63_{\pm 1.58}$	$90.76_{\pm 0.16}$	$18.63_{\pm 2.07}$	$6.71_{\pm 1.38}$
COV+aug	$94.08_{\pm 0.41}$	$8.11_{\pm 2.26}$	$29.03_{\pm 0.72}$	$90.47_{\pm 0.20}$	$13.65_{\pm 1.71}$	$6.78_{\pm 0.09}$
MFD+aug	$93.78_{\pm 0.80}$	$3.87_{\pm 0.84}$	$14.36_{\pm 4.39}$	$89.90_{\pm 0.62}$	$19.36_{\pm 2.41}$	$2.47_{\pm 0.75}$
LBC+aug	$94.39_{\pm 1.44}$	$9.32_{\pm 4.20}$	$36.08_{\pm 11.36}$	$88.66_{\pm 1.25}$	$12.41_{\pm 2.02}$	$2.79_{\pm 1.36}$
SS+CP	$94.54_{\pm 0.09}$	$2.40_{\pm 0.35}$	$37.97_{\pm 2.27}$	$88.70_{\pm 0.82}$	$6.13_{\pm 1.17}$	$4.26_{\pm 1.74}$
RW+CP	$95.19_{\pm 0.13}$	$4.67_{\pm 0.76}$	$25.56_{\pm 2.87}$	$90.87_{\pm 0.28}$	$15.24_{\pm 1.67}$	$6.16_{\pm 0.19}$
COV+CP	$94.29_{\pm 0.18}$	$5.36_{\pm 1.00}$	$51.63_{\pm 0.67}$	$91.23_{\pm 0.37}$	$11.91_{\pm 2.18}$	$6.52_{\pm 1.05}$
MFD+CP	$93.81_{\pm 0.30}$	$3.47_{\pm 0.52}$	$23.31_{\pm 0.74}$	$89.39_{\pm 1.90}$	$15.15_{\pm 1.24}$	$1.90_{\pm 1.03}$
LBC+CP	$95.12_{\pm 0.10}$	$4.72_{\pm 0.87}$	$22.78_{\pm 2.26}$	$89.92_{\pm 0.28}$	$8.33_{\pm 1.07}$	$3.02_{\pm 0.58}$
CKD ($\lambda = 0$)	$94.12_{\pm 0.23}$	$4.31_{\pm 1.47}$	$14.11_{\pm 1.25}$	$90.76_{\pm 0.13}$	$12.42_{\pm 2.69}$	$2.64_{\pm 0.14}$
CKD	$93.08_{\pm 0.46}$	$4.44_{\pm 0.70}$	$13.23_{\pm 1.30}$	$89.26_{\pm 0.45}$	$7.94_{\pm 0.89}$	$1.88_{\pm 0.67}$

Table G.2: **Impact of robustness to G of the teacher model on CKD with Celeb and LFW.** $\boldsymbol{\theta}^T_{\text{CKD}}$ and $\boldsymbol{\theta}^T_{\text{CP}}$ are CKD and CP trained teacher models. $\boldsymbol{\theta}^T_{\text{Scratch}}$ is a vanilla-trained teacher model. RFP^T denotes how a teacher is biased towards G. DEO, CD, RFP are metrics for evaluating GF, CF, and bias towards G, respectively. Since we generate a dataset for RFP measurements on CelebA with "hair length" as G, we report RFP values only for CelebA.

		CelebA				LFW		
Method	RFP^T ↓	Acc ↑	DEO ↓	CD ↓	RFP ↓	Acc ↑	DEO ↓	CD ↓
CKD w/ $\boldsymbol{\theta}^T_{\text{Scratch}}$	15.27	93.08	13.23	4.44	10.85	89.26	1.88	7.94
CKD w/ $\boldsymbol{\theta}^T_{\text{CKD}}$	10.85	93.98	14.37	4.05	11.64	89.17	1.48	8.07
CKD w/ $\boldsymbol{\theta}^T_{\text{CP}}$	20.37	94.25	34.49	3.28	16.61	89.85	9.26	8.32

Table G.3: **The implication of using non-curated IP2P.**

	Acc ↑	CD ↓	DEO ↓
Scratch	78.01	17.90	27.46
CKD	78.49	2.85	7.30
CKD (20%)	79.82	2.77	11.41
CKD (30%)	79.76	2.88	12.78
CKD (40%)	79.70	2.94	12.88
CKD (50%)	79.72	2.85	14.10
CKD (60%)	79.61	3.01	14.94

Table G.4: **The accuracy and DEO on counterfactual samples in CelebA-CF.**

	Acc-CTF ↑	DEO-CTF ↓
Scratch	77.22	15.92
SS	81.43	28.74
RW	79.75	23.61
LBC	78.06	33.95
CP	75.21	62.82
CKD	90.08	21.86

Table G.5: **Evaluation of group fairness (GF) and counterfactual fairness (CF) of fair-training for image classification**. The details are the same as Table 5. "HKD+CP" denotes a model that naively combines Knowledge Distillation [12] with CP [36].

	CIFAR-10B (α=0.8)			CelebA (and CelebA-CF)			LFW (and LFW-CF)		
Method	Acc ↑	DEO ↓	CD ↓	Acc ↑	DEO ↓	CD ↓	Acc ↑	DEO ↓	CD ↓
Scratch	78.36	27.28	17.68	95.53	47.10	10.36	90.85	7.66	18.06
HKD+CP	79.18	16.54	2.40	93.95	33.98	4.40	89.11	3.76	8.78
CKD ($\lambda = 0$)	75.63	6.33	8.94	94.12	14.11	4.31	90.76	2.64	12.47
CKD	78.46	7.11	2.86	93.08	13.23	4.44	89.26	1.88	7.94

Table G.6: **CKD on feature vectors vs logits.**

	CIFAR-10B (α=0.8)			LFW		
Method	Acc ↑	DEO ↓	CD ↓	Acc ↑	DEO ↓	CD ↓
CKD w/ logit	78.46	7.11	2.86	89.26	1.88	7.94
CKD w/ feature	78.24	2.64	1.23	88.37	3.98	6.22

Table G.7: **Sensitivity of μ.**

	CelebA				LFW		
μ	Acc ↑	DEO ↓	CD ↓	μ	Acc ↑	DEO ↓	CD ↓
0.01	94.13	14.83	4.4	0.01	90.05	2.82	15.01
0.1	94.13	14.45	3.03	0.1	90.09	2.39	14.19
1.0	93.63	13.84	4.24	1.0	89.78	2.75	13.66
7.0	93.08	13.23	4.44	10.0	89.23	1.85	13.82
10.0	92.93	13.67	4.10	50.0	88.83	1.8	7.80

H Datasheet for dataset

H.1 Motivation

For what purpose was the dataset created? These datasets were created for evaluating counterfactual fairness in image classifiers. Furthermore, since our datasets contain counterfactual images generated from real-world images, our datasets can be also used for analyzing the relationship between counterfactual and group fairness on image datasets. For more discussion of the motivation behind our datasets, see Section 1.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The datasets were created by the authors of this paper who were affiliated with Seoul National University and NAVER AI LAB.

Who funded the creation of the dataset? Funding was provided by the National Research Foundation of Korea (NRF); Institute of Information & Communications Technology Planning & Evaluation (IITP); and the SNU-Naver Hyperscale AI Center.

H.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The instances represent synthetically generated images and corresponding real-world original images from two popular benchmark facial image datasets, CelebA [25] and LFW [13].

How many instances are there in total (of each type, if appropriate)? CelebA-CF and LFW-CF contain a total of 230 and 144 image pairs of original and counterfactual images, respectively.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? We uniformly sampled a subset of test images in CelebA and LFW to balance the target and group labels (see more details in Section 2). Then, we made our datasets including all possible samples according to our filtering process.

What data does each instance consist of? Each instance contains a pair of original and counterfactual images.

Is there a label or target associated with each instance? Yes, there are 40 binary annotations that originated from CelebA and LFW.

Is any information missing from individual instances? No

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? Yes, instances that correspond to a counterfactual pair are explicitly annotated as such in our dataset. Otherwise, there are no relationships between individual instances.

Are there recommended data splits (e.g., training, development/validation, testing)? No, the dataset is created for the purpose of testing.

Are there any errors, sources of noise, or redundancies in the dataset? Using IP2P as the initial step in constructing our datasets might introduce some noise or errors in the datasets. Refer to Appendix B for further details.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? Yes, it is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals' non-public communications)? No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? Yes, the dataset may cause some anxiety about sex labels. See Section 2 and Appendix B.

Does the dataset identify any subpopulations (e.g., by age, gender)? Yes, our datasets were created after evaluating whether counterfactual samples regarding visually perceived sexual traits were generated correctly or not. This evaluation was conducted by five human annotators. Thus, our datasets contain the identification of visually perceived sexual traits which represent some statistically representative features for each sex. See more discussion in Section 2.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? Yes, our datasets are generated from CelebA and LFW, which are facial datasets collected on the internet.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations;

financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? Yes, we set sex as the sensitive attribute and created our CTF samples with the sensitive attribute flipped.

H.3 Collection Process

How was the data associated with each instance acquired? Our datasets are generated through image editing using IP2P (See Section 2).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? Refer to Section 2 for a complete description of our data generation process.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Original test samples in our datasets were uniformly sampled from the test datasets of CelebA and LFW.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The filtering process for our datasets involved five student annotators who received about 18 USD per hour for their wage.

Over what timeframe was the data collected? Our datasets were generated and evaluated over one month.

Were any ethical review processes conducted (e.g., by an institutional review board)? No

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? No, we initially obtained the data from publicly available sources. Subsequently, we edited the data and filtered the edited one through human annotators.

Were the individuals in question notified about the data collection? Not applicable

Did the individuals in question consent to the collection and use of their data? Not applicable

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? Not applicable

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? Not applicable

H.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? Yes, we filtered our generated datasets with human annotators. Refer to Section 2 for a complete description of our filtering process.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? No, however, raw data can be reproduced by applying IP2P as described in Section 2.

Is the software that was used to preprocess/clean/label the data available? Yes, refer to the Section 2.

H.5 Uses

Has the dataset been used for any tasks already? Yes, we applied our datasets to evaluate CF in image classifiers in Section 3 and 6 and analyze the relationship between CF and GF in Section 5.

Is there a repository that links to any or all papers or systems that use the dataset? We will provide a link to a repository on GitHub that includes references to all papers and systems utilizing the dataset.

What (other) tasks could the dataset be used for? There is no other task where our dataset can be used. The dataset is exclusively designed for evaluating counterfactual fairness in real-world image datasets.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Because our datasets were generated through the image editing technique, IP2P [2], they may contain implicit biases or errors, which are present in the IP2P model [27**?**]. While we have conducted a thorough human filtering and validation process to minimize these issues in our dataset, future users should still be aware of these limitations.

Are there tasks for which the dataset should not be used? The dataset should not be employed for tasks where the limitations discussed in Appendix B could pose critical issues, or for tasks that are not for research purposes.

H.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the datasets will be made publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The dataset will be distributed using tarball on the website. Refer to Appendix C.

When will the dataset be distributed? The datasets will be made publicly available upon acceptance.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The datasets will be distributed under the CC BY 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No

H.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The datasets are hosted, supported, and maintained by the authors.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? The corresponding author can be contacted by the e-mail address which will be listed on the first page of this paper after camera-ready.

Is there an erratum? No

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? No future updates are currently planned. However, we will monitor the GitHub repository for related issues and address any problems that arise.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? Not applicable

Will older versions of the dataset continue to be supported/hosted/maintained? Yes, if the datasets are updated, we will maintain the older versions.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Yes, we make our code and datasets public, and hence others can contribute or extend to our work and datasets.