# Conditional Contrastive Learning for Improving Fairness in Self-Supervised Learning

**Martin Q. Ma**[1], **Yao-Hung Hubert Tsai**[1], **Paul Pu Liang**[1], **Han Zhao**[2], **Kun Zhang**[1,3], **Ruslan Salakhutdinov**[1], **& Louis-Philippe Morency**[1]
[1]Carnegie Mellon University [2]University of Illinois at Urbana-Champaign
[3]Mohamed bin Zayed University of Artificial Intelligence
{qianlim, yaohungt, pliang, kunz1, rsalakhu, morency}@cs.cmu.edu
hanzhao@illinois.edu

## Abstract

Contrastive self-supervised learning (SSL) learns an embedding space that maps similar data pairs closer and dissimilar data pairs farther apart. Despite its success, the fairness aspect of contrastive SSL has been overlooked. Without mitigation, contrastive SSL techniques can incorporate sensitive information such as gender or race and cause potentially unfair predictions on downstream tasks. In this paper, we propose a Conditional Contrastive Learning (CCL) approach to improve the fairness of contrastive SSL methods. Our approach samples positive and negative pairs from distributions conditioning on the sensitive attribute. We show that our approach provably maximizes the conditional mutual information between the learned representations of the positive pairs, and reduces the effect of the sensitive attribute by taking it as the conditional variable. On seven datasets, we empirically demonstrate that the proposed approach achieves state-of-the-art downstream performances compared to unsupervised baselines and significantly improves the fairness of contrastive SSL models on multiple fairness metrics.

## 1 Introduction

Contrastive self-supervised learning (contrastive SSL) [6, 18] have performed well in a variety of different vision or language tasks [8, 7, 34]. Such frameworks perform the *contrastive pre-training* by pulling together related data pairs (termed *positive pairs*) and pushing away unrelated pairs (termed *negative pairs*), and then evaluate the learned representation by a *supervised fine-tuning* with labels. They have been extensively studied because of strong empirical results [20, 24].

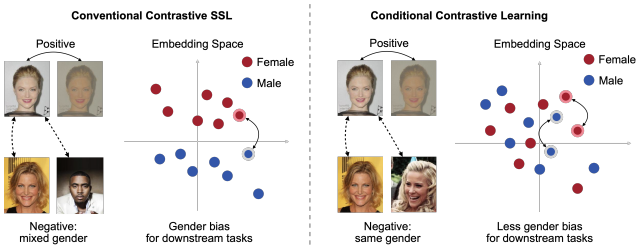


Figure 1: Conventional contrastive SSL vs. the proposed Conditional Contrastive Learning (CCL). Conventional contrastive learning is performed on mixed female and male samples, and the model can easily contrast females from males, creating gender biases. The proposed Conditional Contrastive Learning only samples positive and negative pairs from the same gender, making it harder for the model to pick up gender information and create gender biases.

However, despite the growing popularity of contrastive SSL, the potential issue of fairness in these learned representations has been understudied: **do contrastive SSL models learn fair representations, and how to mitigate potential biases?** We are particularly interested in the scenario where a potential *sensitive attribute*, such as gender or race, is already available in the dataset. We show that without care, contrastive models can incorporate information from the sensitive attributes and cause unfair predictions in downstream tasks. For example, Figure 1 illustrates a case where contrastive SSL is used to learn representations of human faces. A conventional contrastive SSL setup [6] uses two augmented views of the same image as a

Preprint. Under review.

positive pair, and selects random images as negative pairs. In this setup, two images for the positive pair will always share the same gender since both are augmented from the same image, but two images for the negative pairs may have different genders. Therefore, a contrastive SSL model can learn to use gender-related visual attributes to push away mixed-gender images present in negative pairs. By capturing this gender-related information in the embedding space of the learned representation, the representation can potentially causes unfair predictions when it is applied to downstream tasks.

In this paper, we empirically study this potential issue of fairness with contrastive SSL approaches, and propose a new method, Conditional Contrastive Learning (CCL), to reduce the effect from a sensitive attribute during the contrastive pre-training. We focus on scenarios where the sensitive attribute is known and our goal is to mitigate its effect. The proposed CCL approach first defines the sensitive attribute (e.g., gender) as a conditional variable and samples the positive and negative pairs from distributions conditioning on the sensitive attribute. Empirically, this can be efficiently implemented by sampling from the same sensitive attribute, i.e., from the same gender. This simple but effective approach makes it harder for the model to leverage information from the sensitive attribute to distinguish positive pairs from negative pairs. We then proved that the proposed CCL maximizes a lower bound of *conditional* mutual information between the learned representations, which explicitly excludes information from the conditional variable.

We evaluate our approach on five fairness datasets: Adult [11], German [11], COMPAS [1], Crime [11], and Law School [40], and two real-world facial datasets, CelebA [27] and UTK-Face [44]. We study the fairness of contrastive SSL with multiple fairness metrics including demographic parity, equalized odds, and equality of opportunity, with the goal of maintaining strong downstream task performances compared to unsupervised baselines.

## 2 Method

We use $x$ and $y$ to denote the learned representations after encoding two data views $v_1$ and $v_2$ created by stochastic augmentation on images, where $v_1$ and $v_2$ could be a positive (from the same image) or a negative pair (from different images): $x = \text{encoder}(v_1), y = \text{encoder}(v_2)$. Recent work [33, 3, 41, 2, 38] has shown that the successes of contrastive SSL objectives are related to maximizing a lower bound of mutual information shared between the representations of positive pairs. Formally, the conventional contrastive objective, InfoNCE [33], maximizes $\text{MI}(X;Y)$ as follows:

$$\text{InfoNCE} := \sup_f \mathbb{E}_{(x_i,y_i)\sim P_{X,Y}} \left[ \frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{f(x_i,y_i)}}{\frac{1}{n}\sum_{j=1}^{n} e^{f(x_i,y_j)}} \right] \leq D_{\text{KL}}\left(P_{X,Y} \,\|\, P_X P_Y\right) = \text{MI}(X;Y), \tag{1}$$

where the positive pairs $\{(x_i, y_i)\}_{i=1}^n$ are drawn from the joint distribution: $(x_i, y_i) \sim P_{X,Y}$, and the negative pairs $\{(x_i, y_{j\neq i})\}$ are drawn from the product of marginal distributions: $(x_i, y_{j\neq i}) \sim P_X P_Y$. $\text{MI}(X;Y)$ is the mutual information between $X$ and $Y$. The score function is $f(x,y) = \cos\left(g(x), g(y)\right)/\tau$, where $\tau$ is the temperature hyper-parameter.

### 2.1 Conditional Contrastive Learning

Next, we propose our method, Conditional Contrastive Learning (CCL), which differs from conventional contrastive SSL by taking positive and negative pairs from the distributions conditioning on a sensitive attribute referred as $Z$. Our CCL approach reduces information from the sensitive attribute $Z$ by taking $Z$ as the conditional variable between $X$ and $Y$. We assume $Z$ is readily available in the dataset (e.g., gender, race or age), following previous work on fairness [29, 37]. Now we present the proposed **Conditional Contrastive Learning** objective:

$$\text{CCL} := \sup_f \mathbb{E}_{z\sim P_Z} \left[ \mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}} \left[ \frac{1}{n}\sum_{i=1}^{n} \log \frac{e^{f(x_i,y_i)}}{\frac{1}{n}\sum_{j=1}^{n} e^{f(x_i,y_j)}} \right] \right] \tag{2}$$

where the positive pairs $\{(x_i, y_i)\}_{i=1}^n$ represent samples drawn from the *conditional* joint distribution: $(x_i, y_i) \sim P_{X,Y|z}$, while the negative pairs $\{(x_i, y_{j\neq i})\}$ represent samples drawn from the product of *conditional* marginal distributions: $(x_i, y_{j\neq i}) \sim P_{X|z}P_{Y|z}$. The score function is $f(x,y) = \cos\left(g(x), g(y)\right)/\tau$, same as (1). The difference between CCL and InfoNCE can be phrased as follows: CCL first samples $z \sim Z$, and then samples positive and negative pairs from $P_{X,Y|z}$ and $P_{X|z}P_{Y|z}$, respectively. InfoNCE, on the contrary, directly samples from $P_{X,Y}$ and $P_X P_Y$. Sampling from conditional distributions in CCL means all positive and negative pairs share the same outcome $z$ of the sensitive attribute $Z$. Maximizing CCL results in maximizing a lower bound of $\text{CMI}(X;Y|Z)$ between the representation $X$ of data view $V_1$ and representation $Y$ of data view $V_2$ given $Z$. Previous

work [19, 33, 38] has shown that maximizing the information shared between $X$ and $Y$ can produce a good embedding space that has high *representation quality* for downstream tasks.

### 2.2 Theoretical Motivation

In this section, we provide the theoretical motivation for our CCL method. In particular, we are interested in understanding why our method can reduce the information related to the sensitive attribute $Z$. We defer detailed proofs to Appendix. Recall that InfoNCE is maximizing a lower bound of $D_{\mathrm{KL}}\left(P_{X,Y} \,\|\, P_X P_Y\right)$ as shown in (1). Similarly, our CCL method aims to maximize the divergence between $P_{XY|z}$ and $P_{X|z}P_{Y|z}$ for all $z \sim P_Z$, leading to a connection with conditional mutual information $\mathrm{MI}(X;Y|Z)$. First, we define conditional mutual information (CMI):

$$\mathrm{CMI}(X;Y|Z) := \mathbb{E}_{z\sim Z}\left[D_{\mathrm{KL}}\left(P_{X,Y|Z=z} \| P_{X|Z=z}P_{Y|Z=z}\right)\right], = \int_{\mathcal{Z}} D_{\mathrm{KL}}\left(P_{X,Y|Z} \| P_{X|Z}P_{Y|Z}\right) \mathrm{d}P_Z, \tag{3}$$

which measures the expected mutual information of $X$ and $Y$ given $Z$. Intuitively, $\mathrm{CMI}(X;Y|Z)$ measures the averaged shared information by $X$ and $Y$ but exclude the effect from $Z$ [28], as shown in Figure 2. This is because conditioning $Z = z$ means taking $Z = z$ as known and, therefore, ignoring the effect of $Z$ [32]. By ignoring the effect of $Z$, $\mathrm{CMI}(X;Y|Z)$ explicitly excludes the information from $Z$ when measuring the shared information between $X$ and $Y$. Next, we show our main theoretical result, that the proposed CCL objective is a lower bound of the conditional mutual information $\mathrm{CMI}(X;Y|Z)$:

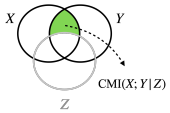


Figure 2: Venn diagram of $\mathrm{CMI}(X;Y|Z)$ in green.

$$\mathrm{CCL} \leq D_{\mathrm{KL}}\left(P_{X,Y} \,\|\, \mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]\right) = \text{Weak-CMI}\left(X;Y|Z\right) \leq \mathrm{CMI}(X;Y|Z), \tag{4}$$

where Weak-CMI $\left(X;Y|Z\right)$ is the KL-divergence between $P_{X,Y}$ and $\mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]$. This notion can achieve the so-called weak-conditional independence [10, 13, 14]. Weak-CMI $\left(X;Y|Z\right)$ is a lower bound of $\mathrm{CMI}(X;Y|Z)$ and can be seen as capturing only part of information in $\mathrm{CMI}(X;Y|Z)$. More discussions can be found in Appendix.

## 3 Experiments

Table 1: Accuracies and fairness results on five fairness datasets. CCL has better downstream accuracy than existing baselines in four datasets, and exhibits better fairness measurements in 11 out of 18 results.

| | Model | Accuracy (%) (↑) | $\Delta_{DP}$ (↓) | $\Delta_{EO}$ (↓) | $\Delta_{EO_{PP}}$ (↓) |
|---|---|---|---|---|---|
| ADULT | – LAFTR [29] | 84.0 | 0.163 | **0.030** | **0.026** |
| | – Ragonesi et al. [35] | 85.0 | - | **0.030** | - |
| | – DTM [25] | 71.6 | - | 0.050 | - |
| | – FNF [4] | 80.0 | **0.110** | - | - |
| | – SimCLR [6] | 83.1 | 0.210 | 0.410 | 0.320 |
| | – FairMixRep [5] | 85.0 | 0.172 | - | - |
| | – **CCL (Ours)** | **85.4** | **0.110** | 0.070 | 0.040 |
| COMPAS | – DTM [25] | 66.0 | - | 0.200 | - |
| | – FNF [4] | 65.0 | 0.240 | - | - |
| | – SimCLR [6] | **71.2** | 0.103 | 0.227 | 0.134 |
| | – **CCL (Ours)** | 71.0 | **0.080** | **0.132** | **0.081** |
| CRIME | – LCIFR [36] | **84.4** | 0.443 | 0.314 | 0.212 |
| | – FNF [4] | 82.5 | 0.540 | - | - |
| | – SimCLR [6] | 82.1 | 0.502 | 0.530 | 0.383 |
| | – **CCL (Ours)** | 82.6 | **0.211** | **0.224** | **0.183** |
| GERMAN | – LCIFR [36] | 73.1 | 0.102 | 0.080 | 0.063 |
| | – Ragonesi et al. [35] | 74.0 | - | **0.060** | - |
| | – FairMixRep [5] | 71.8 | 0.089 | - | - |
| | – SimCLR [6] | 72.5 | 0.250 | 0.382 | 0.195 |
| | – **CCL (Ours)** | **74.3** | **0.083** | 0.128 | **0.062** |
| LAW | – LCIFR [36] | 84.4 | 0.110 | 0.180 | 0.070 |
| | – FNF [4] | 84.6 | **0.050** | - | - |
| | – SimCLR [6] | 83.6 | 0.086 | 0.212 | 0.110 |
| | – **CCL (Ours)** | **84.8** | 0.051 | **0.153** | **0.056** |

We evaluate the proposed Conditional Contrastive Learning on five fairness datasets: Adult [11], Compas [1], Crime [11], German [11], and Law School [40], and two facial datasets: CelebA [26] and UTKFace [44]. We evaluate on prediction accuracy (all tasks being binary predictions) and three fairness metrics (in the form of distance): Demographic Parity ($\Delta_{DP}$), Equalized Odds ($\Delta_{EO}$), and Equality of Opportunity ( $\Delta_{EO_{PP}}$). Lower metrics suggest better fairness results. Details of the sensitive attributes for each dataset, the differences of the metrics, baselines, and implementation details are in Appendix.

**Fairness datasets' results.** Table 1 shows the results on accuracy and fairness metrics. First, we observe that self-supervised SimCLR and CCL have strong downstream prediction results close to or better than the state-of-the-art baselines in Adult, Compas, German and the Law School datasets. Next, looking at the fairness measurements, we observe that the SimCLR baseline performs significantly worse than unsupervised fairness baselines, sometimes two to three times higher ($\Delta_{DP}$ in Adult, $\Delta_{EO}$ in German, and $\Delta_{EO_{PP}}$ in Adult), confirming our earlier concern that contrastive self-supervised learning will produce highly unfair predictions without bias mitigation. The proposed CCL is much better than SimCLR and very competitive compared to other unsupervised baselines, in terms of fairness criteria: the average improvement over five datasets from SimCLR to CCL is 12.28% on $\Delta_{DP}$, 21.08% on $\Delta_{EO}$, and 13.43% on $\Delta_{EO_{PP}}$.

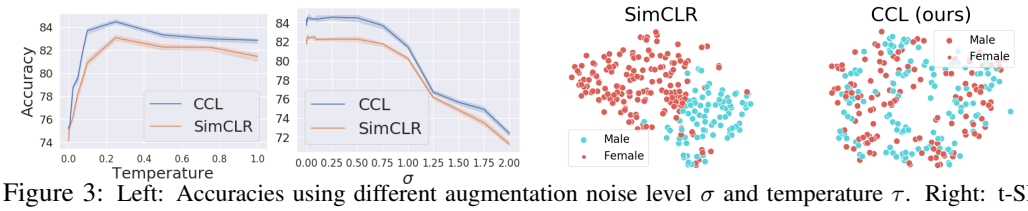


Figure 3: Left: Accuracies using different augmentation noise level $\sigma$ and temperature $\tau$. Right: t-SNE embeddings of conventional contrastive SSL (SimCLR) vs. the proposed Conditional Contrastive Learning.

**Effect of hyper-parameters on downstream performances.** We study two important hyper-parameters using the Adult dataset: the $\sigma$ of the Gaussian noise for data augmentation, and the temperature $\tau$ in Equation 2. The Gaussian noise controls the level of data augmentation, and $\tau$ smooths the distribution of the score output of the encoder. Details can be found in Appendix. We use $\tau \in [0.001, 1]$; and $\sigma \in [0.001, 2]$. The results are shown in Figure 3. Concisely, a mid-range temperature ($\tau = 0.25$) and a mild noise augmentation ($\sigma = 0.25$) to the tabular data help the most in representation learning.

Table 2: Accuracies and fairness results on two vision datasets and four tasks. CCL has better downstream accuracy in all four tasks, and exhibits better or close-to-the-best fairness measurements in 11 out of 12 results.

| | Model | Accuracy (%) (↑) | $\Delta_{DP}$ (↓) | $\Delta_{EO}$ (↓) | $\Delta_{EO_{PP}}$ (↓) |
|---|---|---|---|---|---|
| CELEBA ATTRACTIVE | – MFD [21] | 80.2 | - | **0.050** | - |
| | – Balunovic et al. [4] | 79.4 | - | 0.238 | - |
| | – Morales et al. [30] | 77.7 | - | 0.070 | - |
| | – SimCLR [6] | 81.7 | 0.277 | 0.212 | 0.110 |
| | – RCL [15] | 81.83 | 0.231 | 0.144 | 0.072 |
| | – **CCL (Ours)** | **82.1** | **0.202** | 0.101 | **0.048** |
| CELEBA WAVY | – FactorVAE [22] | 64.5 | - | 0.388 | 0.288 |
| | – FFVAE [9] | 61.0 | - | 0.211 | 0.154 |
| | – SimCLR [6] | **67.7** | 0.403 | 0.355 | 0.210 |
| | – **CCL (Ours)** | **67.7** | **0.202** | **0.189** | **0.102** |
| CELEBA SMILE | – Morales et al. [30] | 88.4 | - | **0.060** | - |
| | – SimCLR [6] | 89.3 | 0.102 | 0.142 | 0.078 |
| | – **CCL (Ours)** | **89.7** | **0.086** | **0.060** | **0.053** |
| UTK GENDER | – AD [42] | 74.7 | - | 0.204 | - |
| | – MFD [21] | 74.7 | - | 0.178 | - |
| | – SimCLR [6] | 78.0 | 0.335 | 0.421 | 0.287 |
| | – **CCL (Ours)** | **78.5** | **0.191** | **0.156** | **0.089** |

**Effect of hyper-parameters on fairness.** We also study the effect of $\sigma$ and $\tau$ in terms of fairness criteria: $\Delta_{DP}$, $\Delta_{EO}$, and $\Delta_{EO_{PP}}$. Overall, a similar trend occurs for three criteria: a large noise ($\sigma > 0.75$) and a large temperature ($\tau > 0.5$) generates the worst representation in terms of fairness metrics ( $\Delta_{DP} > 0.2$, $\Delta_{EO} > 0.3$ and $\Delta_{EO_{PP}} > 0.3$). On the other hand, a large noise ($\sigma > 0.75$) and a medium-to-small temperature ($\tau < 0.5$) generates the best results on fairness, but in these cases the representation performs badly on downstream tasks. Concisely, the results suggest that a larger noise and a mid-range temperature may help remove bias information.

**Vision datasets' results.** We also provide results on the CelebA and UTK datasets in Table 2. We observe that the proposed CCL outperforms unsupervised or self-supervised baselines on the prediction tasks. The CCL also achieves much better fairness criteria than the SimCLR baseline, producing much lower $\Delta_{DP}$, $\Delta_{EO}$, and $\Delta_{EO_{PP}}$ in all four tasks across two datasets.

**CCL embeddings are hard to separate by gender.** We plot the embedding spaces of SimCLR and CCL using t-SNE [39], to verify our claim in the introduction that contrastive embeddings have gender biases. The visualization is in Figure 3. The embeddings of two genders are clearly separated in SimCLR, making it easy for the models to pick up gender bias. On the other hand, the model trained by CCL is much hard to separate two groups, because the sampling from the same gender makes it hard for models to leverage gender bias during contrastive pre-training.

## 4 Conclusion

We introduce Conditional Contrastive Learning (CCL) which perform conditional sampling on the sensitive attribute to remove its effect and improve fairness. By conditioning on the sensitive attribute, CCL samples the positive and negative pairs from the same subgroup, making it harder for the model to leverage gender information. We show CCL is a lower bound of conditional mutual information. CCL significantly improves the fairness of conventional contrastive models, while achieving SOTA downstream performances compared to both contrastive SSL and other unsupervised baselines. We acknowledge that while our method improve fairness on the chosen attributes, it does not provide any guarantees that the resulting method is more or less fair for the other attributes.

## References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23(2016):139–159, 2016.

[2] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019.

[4] Mislav Balunovic, Anian Ruoss, and Martin Vechev. Fair normalizing flows. In *International Conference on Learning Representations*, 2021.

[5] Souradip Chakraborty, Ekansh Verma, Saswata Sahoo, and Jyotishka Datta. Fairmixrep: Self-supervised robust representation learning for heterogeneous data with fairness constraints. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 458–463. IEEE, 2020.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[9] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

[10] JJ Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980.

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[13] Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

[14] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007.

[15] Songwei Ge, Shlok Mishra, Chun-Liang Li, Haohan Wang, and David Jacobs. Robust contrastive learning using negative samples with diminished semantics. *Advances in Neural Information Processing Systems*, 34:27356–27368, 2021.

[16] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[20] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2021.

[21] Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12115–12124, 2021.

[22] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 2020.

[25] Joshua Lee, Yuheng Bu, Prasanna Sattigeri, Rameswar Panda, Gregory Wornell, Leonid Karlinsky, and Rogerio Feris. A maximal correlation approach to imposing fairness in machine learning. *arXiv preprint arXiv:2012.15259*, 2020.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[28] David JC MacKay, David JC Mac Kay, et al. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[29] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[30] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020.

[31] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[32] Jana Novovičová, Petr Somol, Michal Haindl, and Pavel Pudil. Conditional mutual information based feature selection for classification task. In *Iberoamerican Congress on Pattern Recognition*, pages 417–426. Springer, 2007.

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[35] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning unbiased representations via mutual information backpropagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2729–2738, 2021.

[36] Anian Ruoss, Mislav Balunović, Marc Fischer, and Martin Vechev. Learning certified individually fair representations. *arXiv preprint arXiv:2002.10312*, 2020.

[37] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173. PMLR, 2019.

[38] Yao-Hung Hubert Tsai, Yue Wu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Self-supervised learning from a multi-view perspective. In *ICLR*, 2021.

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[40] Linda F Wightman. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[41] Mike Wu, Chengxu Zhuang, Milan Mosse, Daniel Yamins, and Noah Goodman. On mutual information in contrastive learning for visual representations. *arXiv preprint arXiv:2005.13149*, 2020.

[42] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[43] Qinyi Zhang, SL Filippi, Seth Flaxman, and Dino Sejdinovic. Feature-to-feature regression for a two-step conditional independence test. 2017.

[44] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

# A Theoretical Analysis

This section provides the theoretical analysis of Equations (4) and (5) in the main text. The full set of assumptions of all theoretical results and complete proofs of all theoretical results are presented below.

## A.1 Useful lemmas

We first present the following lemmas, which will be later used in the proof:

**Lemma 1** (Nguyen et al. [31] with two variables)**.** *Let $\mathcal{X}$ and $\mathcal{Y}$ be the sample spaces for $X$ and $Y$, $f$ be any function: $(\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, and $\mathcal{P}$ and $\mathcal{Q}$ be the probability measures on $\mathcal{X} \times \mathcal{Y}$. Then,*

$$D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right) = \sup_{f} \mathbb{E}_{(x,y)\sim\mathcal{P}}[f(x,y)] - \mathbb{E}_{(x,y)\sim\mathcal{Q}}[e^{f(x,y)}] + 1.$$

*Proof.* The second-order functional derivative of the objective is $-e^{f(x,y)} \cdot d\mathcal{Q}$, which is always negative. The negative second-order functional derivative implies the objective has a supreme value. Then, take the first-order functional derivative and set it to zero:

$$d\mathcal{P} - e^{f(x,y)} \cdot d\mathcal{Q} = 0.$$

We then get the optimal $f^*(x,y) = \log \frac{d\mathcal{P}}{d\mathcal{Q}}$. Plug in $f^*(x,y)$ into the objective, we obtain

$$\mathbb{E}_{\mathcal{P}}[f^*(x,y)] - \mathbb{E}_{\mathcal{Q}}[e^{f^*(x,y)}] + 1 = \mathbb{E}_{\mathcal{P}}[\log \frac{d\mathcal{P}}{d\mathcal{Q}}] = D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right).$$

□

**Lemma 2** (Nguyen et al. [31] with three variables)**.** *Let $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$ be the sample spaces for $X$, $Y$, and $Y$, $f$ be any function: $(\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}) \to \mathbb{R}$, and $\mathcal{P}$ and $\mathcal{Q}$ be the probability measures on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ . Then,*

$$D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right) = \sup_{f} \mathbb{E}_{(x,y,z)\sim\mathcal{P}}[f(x,y,z)] - \mathbb{E}_{(x,y,z)\sim\mathcal{Q}}[e^{f(x,y,z)}] + 1.$$

*Proof.* The second-order functional derivative of the objective is $-e^{f(x,y,z)} \cdot d\mathcal{Q}$, which is always negative. The negative second-order functional derivative implies the objective has a supreme value. Then, take the first-order functional derivative and set it to zero:

$$d\mathcal{P} - e^{f(x,y,z)} \cdot d\mathcal{Q} = 0.$$

We then get the optimal $f^*(x,y,z) = \log \frac{d\mathcal{P}}{d\mathcal{Q}}$. Plug in $f^*(x,y,z)$ into the objective, we obtain

$$\mathbb{E}_{\mathcal{P}}[f^*(x,y,z)] - \mathbb{E}_{\mathcal{Q}}[e^{f^*(x,y,z)}] + 1 = \mathbb{E}_{\mathcal{P}}[\log \frac{d\mathcal{P}}{d\mathcal{Q}}] = D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right).$$

□

### A.1.1 Immediate results following Lemma 1

**Lemma 3.**

$$\begin{aligned} \textit{Weak-CMI}\left(X;Y|Z\right) &= D_{\mathrm{KL}}\left(P_{X,Y} \,\|\, \mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]\right) \\ &= \sup_{f} \mathbb{E}_{(x,y)\sim P_{X,Y}}[f(x,y)] - \mathbb{E}_{(x,y)\sim \mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]}[e^{f(x,y)}] + 1. \end{aligned}$$

*Proof.* Let $\mathcal{P}$ be $P_{X,Y}$ and $\mathcal{Q}$ be $\mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]$ in Lemma 1. □

**Lemma 4.** $\sup_{f} \mathbb{E}_{(x,y_1)\sim\mathcal{P},(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}} \left[\log \frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^{n} e^{f(x,y_j)}}\right] \leq D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right).$

*Proof.* $\forall f$, we have

$$
\begin{aligned}
D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right) &= \mathbb{E}_{(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right)\right] \\
&\geq \mathbb{E}_{(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[\mathbb{E}_{(x,y_1)\sim\mathcal{P}}\left[\log\frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^{n}e^{f(x,y_j)}}\right] - \mathbb{E}_{(x,y_1)\sim\mathcal{Q}}\left[\frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^{n}e^{f(x,y_j)}}\right] + 1\right] \\
&= \mathbb{E}_{(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[\mathbb{E}_{(x,y_1)\sim\mathcal{P}}\left[\log\frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^{n}e^{f(x,y_j)}}\right] - 1 + 1\right] \\
&= \mathbb{E}_{(x,y_1)\sim\mathcal{P},(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[\log\frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^{n}e^{f(x,y_j)}}\right].
\end{aligned}
$$

The first line comes from the fact that $D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right)$ is a constant. The second line comes from Lemma 1. The third line comes from the fact that $(x, y_1)$ and $(x, y_{2:n})$ are interchangeable when they are all sampled from $\mathcal{Q}$.

To conclude, since the inequality works for all $f$, and hence

$$
\sup_f \mathbb{E}_{(x,y_1)\sim\mathcal{P},(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[\log\frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^{n}e^{f(x,y_j)}}\right] \leq D_{\mathrm{KL}}\left(\mathcal{P} \,\|\, \mathcal{Q}\right).
$$

□

Note that Lemma 4 does not require $n \to \infty$, which is a much more practical setting compared to the analysis made only when $n \to \infty$. And a remark is that the equality holds in Lemma 4 when $n \to \infty$.

### A.1.2 Immediate results following Lemma 2

**Lemma 5.**

$$
\begin{aligned}
\mathit{CMI}(X;Y|Z) &= \mathbb{E}_{P_Z}\left[D_{\mathrm{KL}}\left(P_{X,Y|Z} \,\|\, P_{X|Z}P_{Y|Z}\right)\right] \\
&= D_{\mathrm{KL}}\left(P_{X,Y,Z} \,\|\, P_Z P_{X|Z} P_{Y|Z}\right) \\
&= \sup_f \mathbb{E}_{(x,y,z)\sim P_{X,Y,Z}}[f(x,y,z)] - \mathbb{E}_{(x,y,z)\sim P_Z P_{X|Z} P_{Y|Z}}[e^{f(x,y,z)}] + 1.
\end{aligned}
$$

*Proof.* Let $\mathcal{P}$ be $P_{X,Y,Z}$ and $\mathcal{Q}$ be $P_Z P_{X|Z} P_{Y|Z}$ in Lemma 2. □

### A.1.3 Showing Weak-CMI $(X;Y|Z) \leq$ CMI$(X;Y|Z)$

**Proposition 6.** *Weak-CMI* $(X;Y|Z) \leq$ *CMI*$(X;Y|Z)$.

*Proof.* According to Lemma 3,

$$
\begin{aligned}
\text{Weak-CMI}\left(X;Y|Z\right) &= \sup_f \mathbb{E}_{(x,y)\sim P_{X,Y}}[f(x,y)] - \mathbb{E}_{(x,y)\sim \mathbb{E}_{P_Z}[P_{X|Z}P_{Y|Z}]}[e^{f(x,y)}] + 1 \\
&= \sup_f \mathbb{E}_{(x,y,z)\sim P_{X,Y,Z}}[f(x,y)] - \mathbb{E}_{(x,y,z)\sim P_Z P_{X|Z} P_{Y|Z}}[e^{f(x,y)}] + 1.
\end{aligned}
$$

Let $f_1^*(x,y)$ be the function when the equality for Weak-CMI $(X;Y|Z)$ holds, and let $f_2^*(x,y,z) = f_1^*(x,y)$ ($f_2^*(x,y,z)$ will not change $\forall z \sim P_Z$):

$$
\begin{aligned}
\text{Weak-CMI}\left(X;Y|Z\right) &= \mathbb{E}_{(x,y,z)\sim P_{X,Y,Z}}[f_1^*(x,y)] - \mathbb{E}_{(x,y,z)\sim P_Z P_{X|Z} P_{Y|Z}}[e^{f_1^*(x,y)}] + 1 \\
&= \mathbb{E}_{(x,y,z)\sim P_{X,Y,Z}}[f_2^*(x,y,z)] - \mathbb{E}_{(x,y,z)\sim P_Z P_{X|Z} P_{Y|Z}}[e^{f_2^*(x,y,z)}] + 1.
\end{aligned}
$$

Comparing the equation above to Lemma 5,

$$
\text{CMI}(X;Y|Z) = \sup_f \mathbb{E}_{(x,y,z)\sim P_{X,Y,Z}}[f(x,y,z)] - \mathbb{E}_{(x,y,z)\sim P_Z P_{X|Z} P_{Y|Z}}[e^{f(x,y,z)}] + 1,
$$

we conclude Weak-CMI $(X;Y|Z) \leq$ CMI$(X;Y|Z)$. □

### A.2 Proof of a tighter bound of CMI$(X;Y|Z)$

Next, we show a bound of CMI$(X;Y|Z)$ which is tighter than the proposed CCL. We term this bound as Tight-CCL.

**Proposition 7** (A tighter bound of CMI$(X;Y|Z)$)**.**

$$\begin{aligned}\textit{Tight-CCL} &:= \sup_f \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log \frac{e^{f(x_i,y_i,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j,z)}}\right]\right] \\ &\le \mathbb{E}_{P_Z}\left[D_{\mathrm{KL}}\left(P_{X,Y|Z}\,\|\,P_{X|Z}P_{Y|Z}\right)\right] = \textit{CMI}(X;Y|Z),\end{aligned}$$

*Proof.* Given a $z\sim P_Z$, we let $\mathcal{P}=P_{X,Y|Z=z}$ and $\mathcal{Q}=P_{X|Z=z}P_{Y|Z=z}$. Then,

$$\mathbb{E}_{(x,y_1)\sim\mathcal{P},(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[\log\frac{e^{f(x,y_1,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x,y_j,z)}}\right] = \mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log\frac{e^{f(x_i,y_i,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j,z)}}\right].$$

The only variables in the above equation are $X$ and $Y$ with $Z$ being fixed at $z$, and hence the following can be obtained via Lemma 4:

$$\mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log\frac{e^{f(x_i,y_i,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j,z)}}\right] \le D_{\mathrm{KL}}\left(\mathcal{P}\,\|\,\mathcal{Q}\right) = D_{\mathrm{KL}}\left(P_{X,Y|Z=z}\,\|\,P_{X|Z=z}P_{Y|Z=z}\right).$$

The above inequality works for any function $f(\cdot,\cdot,\cdot)$ and any $z\sim P_Z$, and hence

$$\sup_f \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log\frac{e^{f(x_i,y_i,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j,z)}}\right]\right] \le \mathbb{E}_{P_Z}\left[D_{\mathrm{KL}}\left(P_{X,Y|Z}\,\|\,P_{X|Z}P_{Y|Z}\right)\right].$$

□

We discuss the similarities and differences between CCL and Tight-CCL. Both are lower bounds of conditional mutual information CMI$(X;Y|Z)$, and both share formulations similar to InfoNCE [33]. The differences are that the scoring function $f(x,y,z)$ of Tight-CCL takes $z$ as input, while CCL does not. Taking $z$ as input makes Tight-CCL a tighter bound than CCL, which we show in Proposition 9. The reason we do not take $z$ as an input in the CCL is because the sensitive attribute $z$ in our setup is mostly binary, carries little information, and empirically Tight-CCL performs very similar to the proposed CCL (see Section B). CCL, on the other hand, has a simpler formulation and is easier to adapt to existing contrastive frameworks.

### A.3 Proof of Equation (4) in the Main Text

**Proposition 8** (Conditional Contrastive Learning (CCL), restating Equation (4) in the main text)**.**

$$\begin{aligned}\textit{CCL} &:= \sup_f \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log\frac{e^{f(x_i,y_i)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j)}}\right]\right] \\ &\le D_{\mathrm{KL}}\left(P_{X,Y}\,\|\,\mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]\right) = \textit{Weak-CMI}\left(X;Y|Z\right) \le \textit{CMI}(X;Y|Z).\end{aligned}$$

*Proof.* By defining $\mathcal{P}=P_{X,Y}$ and $\mathcal{Q}=\mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]$, we have

$$\mathbb{E}_{(x,y_1)\sim\mathcal{P},(x,y_{2:n})\sim\mathcal{Q}^{\otimes(n-1)}}\left[\log\frac{e^{f(x,y_1)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x,y_j)}}\right] = \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log\frac{e^{f(x_i,y_i)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j)}}\right]\right].$$

Via Lemma 4, we have

$$\sup_f \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim P_{X,Y|z}{}^{\otimes n}}\left[\log\frac{e^{f(x_i,y_i)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j)}}\right]\right] \le D_{\mathrm{KL}}\left(P_{X,Y}\,\|\,\mathbb{E}_{P_Z}\left[P_{X|Z}P_{Y|Z}\right]\right).$$

Combing with Proposition 6 that Weak-CMI $\left(X;Y|Z\right)\le$ CMI$(X;Y|Z)$, we conclude the proof. □

### A.4 Showing CCL is a lower bound of Tight-CCL

**Proposition 9.**

$$\begin{aligned} \textit{CCL} &:= \sup_f \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim {P_{X,Y|z}}^{\otimes n}}\left[\log \frac{e^{f(x_i,y_i)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j)}}\right]\right] \\ \leq \quad \textit{Tight-CCL} &:= \sup_f \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim {P_{X,Y|z}}^{\otimes n}}\left[\log \frac{e^{f(x_i,y_i,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f(x_i,y_j,z)}}\right]\right]. \end{aligned}$$

*Proof.* Let $f_1^*(x,y)$ be the function when the equality holds in CCL, and let $f_2^*(x,y,z) = f_1^*(x,y)$ $\Big($ $f_2^*(x,y,z)$ will not change $\forall z \sim P_Z$ $\Big)$:

$$\text{CCL} := \mathbb{E}_{z\sim P_Z}\left[\mathbb{E}_{(x_i,y_i)\sim {P_{X,Y|z}}^{\otimes n}}\left[\log \frac{e^{f_2^*(x_i,y_i,z)}}{\frac{1}{n}\sum_{j=1}^n e^{f_2^*(x_i,y_j,z)}}\right]\right].$$

Since the equality holds with the supreme function in Tight-CCL, and hence

$$\text{CCL} \leq \text{Tight-CCL}.$$

□

### A.5 Discussions of Weak Conditional Independence

We have the weak conditional independence between $X$ and $Y$ given $Z$ when Weak-CMI $(X;Y|Z) = 0$. First, Weak-CMI $(X;Y|Z) = 0$ is a necessary but not sufficient condition for CMI$(X;Y|Z) = 0$, suggesting that conditional independence implies weak conditional independence. For example, if $X$, $Y$, and $Z$ are pairwise independent but jointly dependent, Weak-CMI $(X;Y|Z) = 0$ but CMI$(X;Y|Z)$ may not be zero. Although weak conditional independence does not fully characterize conditional independence, it has been shown to be widely useful in practice. For instance, testing weak conditional independence can be simpler and more powerful than the original conditional independence test [43]. Our approach benefits from the notion of weak conditional independence in similar ways. Also, we prove that Weak-CMI $(X;Y|Z)$ is a lower bound of CMI$(X;Y|Z)$ and can be seen as a more "conservative" measurement of CMI$(X;Y|Z)$, capturing only part of information in CMI$(X;Y|Z)$.

## B Experimental Details

### B.1 Methodology

We follow the setup from the contrastive SSL learning literature [6, 18], which contains two stages: contrastive pre-training and supervised fine-tuning. We use the SimCLR framework [6]. In contrastive pre-training, we train an encoder without any labels. In the supervised fine-tuning stage, we freeze the encoder and fune-tine an additional small network with the downstream labels. We then evaluate the fine-tuned representations on the test splits of the corresponding datasets. For fairness datasets, we use a three-layer neural network with hidden dimension 100 as the encoder and a linear layer as the fine-tuning network. For vision datasets, we use a ResNet-50 [17] as the encoder and a two-layer network as the fine-tuning network. There are two types of baselines: unsupervised and SSL baselines. Unsupervised baselines include models dedicated for improving fairness in unsupervised representations. The SSL baseline includes implementations of SimCLR. We did not include supervised fair representation models, as they often require labels and sensitive attributes available simultaneously, which is not our case.

### B.2 Fairness Tabular Dataset Details

**UCI Adult** [11] focuses on predicting income of a person exceeds fifty thousand per year based on census data. It has a total of $48,842$ samples, with a pre-determined training split of $32,561$ samples

Table 3: Details of datasets, the chosen sensitive attributes, and the corresponding prediction tasks. There are three prediction task for CelebA: attractiveness, weary hair, and smiling.

| Datasets | Type | Number of Samples | Sensitive Attribute | Prediction Task(s) |
|---|---|---|---|---|
| Adult [11] | Tabular | $48,842$ | Gender | Income level |
| Compas [1] | Tabular | $5,278$ | Race | Recidivism |
| Crime [11] | Tabular | $1,994$ | Race | Crime level |
| German [11] | Tabular | $1,000$ | Age | Credit approval |
| Law School [40] | Tabular | $36,022$ | Race | Exam result |
| UTKFace [44] | Vision | $23,708$ | Race | Age |
| CelebA [26] | Vision | $202,599$ | Gender | Multiple |

and a test split of $16,281$ samples. We choose the gender attribute as the sensitive attribute. It has the CC0: Public Domain License.

**UCI German** [11] focuses on predicting whether a person has good credit or not based on a set of attributes. It has a total of $10,00$ samples. We follow the split in Ruoss et al. [36], where $80\%$ of samples are drawn randomly and used as the training set and $20\%$ samples are drawn randomly and used as the test set. We choose the age attribute as the sensitive attribute, which is determined by whether the individual's age exceeds a threshold. It has the Database Contents License v1.0.

**UCI Crime**: The Communities and Crime dataset [11] contains data including socioeconomic, law enforcement, and crime information for US communities. Specifically, it focuses on predicting whether a specific community is above or below the median number of violent crimes per population. It has $1,994$ samples. We follow the split in Ruoss et al. [36], where $80\%$ of samples are drawn randomly and used as the training set and $20\%$ samples are drawn randomly and used as the test set. We choose the race attribute as the sensitive attribute, which is determined by whether the individual has race white. It has the Database Contents License v1.0.

**COMPAS**: The Recidivism Risk COMPAS Score dataset [1] contains a variety of demographic and crime information collected on the use of the COMPAS risk assessment tool in Broward County, Florida Angwin. It focuses on predicting recidivism (whether a criminal will reoffend or not) in the USA. It has $5,728$ samples. We follow the split in Ruoss et al. [36], where $80\%$ of samples are drawn randomly and used as the training set and $20\%$ samples are drawn randomly and used as the test set. We choose the predefined binary race attribute as the sensitive attribute. It has the Database Contents License v1.0.

**Law School**: The Law School dataset is from the Law School Admission Study [40]. It has application records for 25 different law schools. It focuses on predicting whether a student passes the law school bar exam. It has $36,022$ samples. We follow the split in Ruoss et al. [36], where $80\%$ of samples are drawn randomly and used as the training set and $20\%$ samples are drawn randomly and used as the test set. We choose the race attribute as the sensitive attribute, which is determined by whether the individual has race white. It has the Database Contents License v1.0.

**Dataset pre-processing** : We perform the following types of preprocessing on all five fairness datasets: first, we standardize each numerical feature of the data to zero mean and unit variance. Next, we use one-hot encoding scheme for categorical features. Then, we drop rows and columns with missing values, and lastly we split into train, test and validation sets. For the contrastive pre-training, we augment each data sample using two noise vectors sampled from an isotropic Guassian distribution, where the variance is a hyper-parameter. For the supervised fine-tuning, all downstream classification tasks are binary prediction tasks.

**Personal identifiable information** : Personally identifiable information is not available in all five datasets, because the authors of the datasets explicitly remove personal information when creating the datasets.

### B.3 Differences of fairness metrics

We use three types of fairness metrics: the demographic parity (DP [12]) distance $\Delta_{DP}$ [29], equalized odds (EO [16]) distance $\Delta_{EO}$ [37], and the equality of opportunity ($EO_{PP}$ [16]) distance $\Delta_{EO_{PP}}$ [37].

Given the data $X$, the sensitive attribute $Z$ indicating group information, the ground truth downstream task label $l$, and the label prediction from the model $\hat{l}$, the $\Delta_{DP}$ calculates the expected difference (in absolute value) in model predictions between two groups: $\Delta_{DP} = |\mathbb{P}\{\hat{l} = 1|Z = 0\} - \mathbb{P}\{\hat{l} = 1|Z = 1\}|$. The second metric, $\Delta_{EO}$, calculates the sum of the expected difference (in absolute value) of the True Positive Rate and the False Positive Rate of the model predictions between two groups: $\Delta_{EO} = |\mathbb{P}\{\hat{l} = 1|Z = 0, l = 1\} - \mathbb{P}\{\hat{l} = 1|Z = 1, l = 1\}| + |\mathbb{P}\{\hat{l} = 1|Z = 0, l = 0\} - \mathbb{P}\{\hat{l} = 1|Z = 1, l = 0\}$. As a relaxation of $\Delta_{EO}$, $\Delta_{EO_{PP}}$ calculates the expected difference (in absolute value) of only the True Positive Rate of the model predictions between two groups: $\Delta_{EO_{PP}} = |\mathbb{P}\{\hat{l} = 1|Z = 0, l = 1\} - \mathbb{P}\{\hat{l} = 1|Z = 1, l = 1\}|$. $\Delta_{DP}$, $\Delta_{EO}$, $\Delta_{EO_{PP}}$ range from 0 to 1, and a smaller distance is desirable. $\Delta_{DP} = 0$ corresponds to the statistical independence of the sensitive attribute $Z$ and the prediction $\hat{l}$, and $\Delta_{EO} = 0$ corresponds to the conditional independence of $Z$ and $\hat{l}$ given the true label $l$. $\Delta_{DP} = 0$ suggests that members of different groups (e.g., female and male) have the same chance of receiving a favorable prediction ($l = 1$).

### B.4 Fairness tabular dataset training details and results

The self-supervised baseline, SimCLR [33], and the unsupervised baseline, LCIFR [36] are re-implemented based on Ruoss et al. [36]. To stochastically augment tabular features (e.g., age, education, occupation, etc in Adult [11] dataset) and create data views similar to Chen et al. [6], we first standardize each tabular feature, and then use noise vectors from an isotropic Gaussian to perturb the features. Each dataset uses one separate Gaussian, and $\sigma$ of the Gaussian is treated as a hyper-parameter for different datasets. Then we feed the augmented views to the encoder, and then use the output of the encoder to estimate our proposed CCL.

We follow the implementation from [36]. We use a three-layer neural network with hidden dimension 100 as the encoder and a linear layer as the fine-tuning network. We train $100$ epochs and report the result. For pre-training, we use the Adam [23] optimizer, with a batch size of $256$, a learning rate of $0.001$, and a weight decay of $0.01$. For fine-tuning, we use the same optimizer, batch size, weight decay, but a slightly larger learning rate $0.005$.

**Results.** We include the results in Table 4. All entries with $-$ indicate that the corresponding metrics are not reported in the original papers. The following results are read off from the figures in the paper: LAFTR [29], DTM [25], and FNF [4]. We include the confidence intervals of the results, and bold the entries that have overlapping confidence intervals with the best performing entries in that dataset. SimCLR is a re-implementation of Chen et al. [6] on the new datasets. Tight-CCL represents a tighter bound of conditional mutual information, which is introduced and discussed in Proposition 7. From the results, we can conclude that CCL outperforms all baselines on eleven out of the eighteen fairness metrics. Also, CCL outperforms all baselines on downstream accuracy on four out of five datasets. We note that Tight-CCL performs very close to CCL, sometimes better than CCL in terms of fairness metrics. This may be due to that the Tight-CCL is a tighter bound of conditional mutual information, and optimizing Tight-CCL leads to a representation closer to conditional mutual information maximization. Because conditional mutual information explicitly excludes information from the sensitive attribute $Z$, Tight-CCL is able to remove slightly more effect from the sensitive attribute than CCL. We use CCL in the main text as it has a simpler formulation and is easier to adapt to existing contrastive frameworks.

Table 4: Accuracies and fairness results on five fairness datasets with confidence intervals. CCL has better downstream accuracy than existing unsupervised and self-supervised baselines in four datasets, and exhibits better fairness measurements in most cases.

| | Model | Accuracy (%) (↑) | $\Delta_{DP}$ (↓) | $\Delta_{EO}$ (↓) | $\Delta_{EO_{PP}}$ (↓) |
|---|---|---|---|---|---|
| ADULT | **Unsupervised** | | | | |
| | – LAFTR [29] | 84.0 | 0.163 | **0.030** | **0.026** |
| | – Ragonesi et al. [35] | 85.0 | - | **0.030** | - |
| | – DTM [25] | 71.6 | - | 0.050 | - |
| | – FNF [4] | 80.0 | **0.110** | - | - |
| | **Self-Supervised** | | | | |
| | – FairMixRep [5] | 85.0 | 0.172 | - | - |
| | – SimCLR [6] | $83.1 \pm 0.48$ | $0.210 \pm 0.04$ | $0.410 \pm 0.05$ | $0.320 \pm 0.04$ |
| | – **CCL (Ours)** | **85.4** $\pm 0.53$ | **0.110** $\pm 0.02$ | $0.070 \pm 0.01$ | $0.090 \pm 0.01$ |
| | – **Tight-CCL (Ours)** | **85.3** $\pm 0.48$ | **0.108** $\pm 0.03$ | $0.068 \pm 0.01$ | $0.093 \pm 0.01$ |
| COMPAS | **Unsupervised** | | | | |
| | – DTM [25] | 66.0 | - | 0.200 | - |
| | – FNF [4] | 65.0 | 0.240 | - | - |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | **71.2** $\pm 0.33$ | $0.103 \pm 0.01$ | $0.227 \pm 0.06$ | $0.134 \pm 0.04$ |
| | – **CCL (Ours)** | **71.0** $\pm 0.25$ | **0.080** $\pm 0.01$ | **0.132** $\pm 0.03$ | **0.081** $\pm 0.01$ |
| | – **Tight-CCL (Ours)** | **70.8** $\pm 0.28$ | **0.090** $\pm 0.01$ | **0.142** $\pm 0.02$ | **0.101** $\pm 0.02$ |
| CRIME | **Unsupervised** | | | | |
| | – LCIFR [36] | **84.4** | 0.443 | 0.314 | 0.212 |
| | – FNF [4] | 82.5 | 0.540 | - | - |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | $82.1 \pm 0.32$ | $0.502 \pm 0.08$ | $0.530 \pm 0.02$ | $0.383 \pm 0.01$ |
| | – **CCL (Ours)** | $82.6 \pm 0.24$ | **0.211** $\pm 0.02$ | **0.224** $\pm 0.02$ | **0.183** $\pm 0.01$ |
| | – **Tight-CCL (Ours)** | $82.5 \pm 0.30$ | **0.208** $\pm 0.02$ | **0.222** $\pm 0.02$ | **0.181** $\pm 0.02$ |
| GERMAN | **Unsupervised** | | | | |
| | – LCIFR [36] | 73.1 | 0.102 | 0.080 | 0.063 |
| | – Ragonesi et al. [35] | 74.0 | - | **0.060** | - |
| | **Self-Supervised** | | | | |
| | – FairMixRep [5] | 71.8 | 0.089 | - | - |
| | – SimCLR [6] | $72.5 \pm 0.11$ | $0.250 \pm 0.05$ | $0.382 \pm 0.06$ | $0.195 \pm 0.04$ |
| | – **CCL (Ours)** | **74.3** $\pm 0.28$ | **0.083** $\pm 0.01$ | $0.128 \pm 0.03$ | **0.062** $\pm 0.01$ |
| | – **Tight-CCL (Ours)** | **74.4** $\pm 0.25$ | **0.085** $\pm 0.01$ | $0.126 \pm 0.02$ | **0.062** $\pm 0.01$ |
| LAW SCHOOL | **Unsupervised** | | | | |
| | – LCIFR [36] | 84.4 | 0.110 | 0.180 | 0.070 |
| | – FNF [4] | 84.6 | **0.050** | - | - |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | $83.6 \pm 0.54$ | $0.086 \pm 0.02$ | $0.212 \pm 0.04$ | $0.110 \pm 0.02$ |
| | – **CCL (Ours)** | **84.8** $\pm 0.50$ | **0.051** $\pm 0.01$ | **0.153** $\pm 0.03$ | **0.056** $\pm 0.01$ |
| | – **Tight-CCL (Ours)** | **84.5** $\pm 0.44$ | **0.050** $\pm 0.01$ | **0.150** $\pm 0.02$ | **0.055** $\pm 0.01$ |

### B.5 Effect of hyper-parameters on downstream performance

FOr representation learning, we observe that a mid-range temperature $\tau = 0.25$ achieves the best results. The prediction accuracy begins to increase drastically as $\tau$ goes from $0.001$, tops at $\tau = 0.25$, and start decreasing slightly from $\tau = 0.5$. Next, for the noise level, we observe that a small $\sigma$ ranging from $0.001$ to $0.75$ achieves similar results (around or above $84\%$), peaks at $\sigma = 0.25$, and then degrades fast after $0.75$.

**Computational resource** We perform all experiments on a single GeForce RTX 2080 Ti GPU and a 32-core Intel CPU processor. Training 100 epochs in different datasets vary based on the size of the dataset, but the overall training time of 100 epochs on one dataset is below an hour.

### B.6 Vision dataset details

**CelebA** [27] is a human facial recognition dataset that contains more than $200,000$ images of celebrity faces, where each facial image is annotated with 40 human-labeled binary attributes, including gender. Among the attributes, we select attractive, smile, and wavy hair and use them to form three separate binary classification tasks. The sensitive attribute is gender. The license of CelebA dataset claims that it is available for noncommercial research purposes only.

**UTKFace** [44] is a human facial recognition dataset that contains more than $20,000$ images of human faces in a variety of age groups and races, where each facial image is annotated with three human-labeled binary attributes, including age, gender, and ethnicity. Among the attributes, we select age as the binary classification task (if the age of the individual is above a threshold). The sensitive attribute is the race attribute. The license of UTKFace dataset claims that it is available for noncommercial research purposes only.

**Dataset pre-processing** : For CelebA, we directly use the pre-defined training and test sets from the PyTorch data loader for CelebA. For UTKFace, we use a random $20\%$ of all samples as the test set. The data augmentation details of both datasets will be included in Section B.7.

**Personal identifiable information** : Personally identifiable information is not available in the UTKFace data set, because the authors of the data sets explicitly remove personal information when creating the data set. For the CelebA dataset, each person has an ID, but the identity is not explicitly revealed (although users can infer the identities of some celebrities). Both datasets contain the annotated information, such as age, gender, and other facial attributes of the individuals in the images.

### B.7 Vision Dataset Training Details and Results

We follow the implementation from [6]. We use a ResNet-50 as the encoder and a two-layer network with hidden dimension $512$ as the fine-tuning network. We train $100$ epochs and report the result. For the contrastive pre-training, we use the Adam [23] optimizer, with a batch size of $256$, a learning rate of $0.0003$, and a weight decay of $10^{-6}$. For the supervised fine-tuning, we use the same optimizer, batch size, weight decay, but a slightly larger learning rate $0.001$.

**Results.** We include the results in Table 6. All entries with $-$ indicate that the corresponding metrics are not reported in the original papers. We include the confidence intervals of the results, and bold the entries that have overlapping confidence intervals with the best performing entries in that dataset. Similar to our observation in Section B.4, from the results we can conclude that CCL outperforms all baselines on eleven out of the twelve fairness metrics. Also, CCL outperforms all baselines on downstream accuracy on all four tasks. Tight-CCL also performs very close to CCL, although some downstream task performances of Tight-CCL is slightly worse than that of CCL.

**Computational resource** We perform all experiments on a single GeForce RTX 2080 Ti GPU and a 32-core Intel CPU processor. Training $100$ epochs on CelebA or UTKFace takes approximately $20-24$ hours, depending on the server's condition.

### B.8 Contrastive SSL performs worse on fairness than supervised methods

To study whether contrastive SSL models perform better or worse on fairness criteria than a supervised counterpart, we train two ResNet-18 models [17], one with contrastive pre-training then fine-tuning [18], and one with supervised training. Both models have the same architecture and training hyperparameters. From Table 5, given similar performance, contrastive SSL has significantly larger fairness differences, suggesting that contrastive SSL can produce downstream predictions that perform much worse on fairness criteria than its supervised counterpart.

Table 5: Fairness criteria on contrastive SSL vs. supervised counterpart. Contrastive SSL has mugh higher level of fairness differences based on the three metrics.

| | Accuracy | $\Delta_{DP}$ (↓) | $\Delta_{EO}$ (↓) | $\Delta_{EO_{PP}}$ (↓) |
|---|---|---|---|---|
| Supervised | 80.4 | 0.214 | 0.186 | 0.080 |
| Contrastive SSL | 80.1 | 0.355 | 0.541 | 0.310 |

Table 6: Accuracies and fairness results on two vision datasets on four prediction tasks with confidence intervals. Best results are bold. CCL has better downstream accuracy in all four tasks, and exhibits better or close-to-the-best fairness measurements in 11 out of 12 results.

| | Model | Accuracy (%) (↑) | $\Delta_{DP}$ (↓) | $\Delta_{EO}$ (↓) | $\Delta_{EO_{PP}}$ (↓) |
|---|---|---|---|---|---|
| CELEBA ATTRACTIVE | **Unsupervised** | | | | |
| | – MFD [21] | 80.2 | - | **0.050** | - |
| | – Balunovic et al. [4] | 79.4 | - | 0.238 | - |
| | – Morales et al. [30] | 77.7 | - | 0.070 | - |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | **81.7** $\pm$ 0.32 | 0.277 $\pm$ 0.04 | 0.212 $\pm$ 0.03 | 0.110 $\pm$ 0.01 |
| | – **CCL (Ours)** | **82.1** $\pm$ 0.24 | **0.202** $\pm$ 0.03 | 0.101 $\pm$ 0.01 | **0.048** $\pm$ 0.01 |
| | – **Tight-CCL (Ours)** | **81.9** $\pm$ 0.33 | **0.200** $\pm$ 0.02 | 0.106 $\pm$ 0.02 | **0.052** $\pm$ 0.01 |
| CELEBA WAVY HAIR | **Unsupervised** | | | | |
| | – FactorVAE [22] | 64.5 | - | 0.388 | 0.288 |
| | – FFVAE [9] | 61.0 | - | 0.211 | 0.154 |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | **67.7** $\pm$ 0.76 | 0.403 $\pm$ 0.05 | 0.355 $\pm$ 0.04 | 0.210 $\pm$ 0.02 |
| | – **CCL (Ours)** | **67.7** $\pm$ 0.69 | **0.202** $\pm$ 0.02 | **0.189** $\pm$ 0.02 | **0.102** $\pm$ 0.01 |
| | – **Tight-CCL (Ours)** | **67.8** $\pm$ 0.44 | **0.198** $\pm$ 0.03 | **0.172** $\pm$ 0.02 | **0.093** $\pm$ 0.01 |
| CELEBA SMILE | **Unsupervised** | | | | |
| | – Morales et al. [30] | 88.4 | - | **0.060** | - |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | **89.3** $\pm$ 0.33 | 0.102 $\pm$ 0.01 | 0.142 $\pm$ 0.01 | 0.078 $\pm$ 0.01 |
| | – **CCL (Ours)** | **89.7** $\pm$ 0.25 | **0.086** $\pm$ 0.01 | **0.060** $\pm$ 0.01 | **0.053** $\pm$ 0.01 |
| | – **Tight-CCL (Ours)** | **89.5** $\pm$ 0.27 | **0.084** $\pm$ 0.01 | **0.060** $\pm$ 0.01 | **0.056** $\pm$ 0.01 |
| UTKFACE GENDER | **Unsupervised** | | | | |
| | – AD [42] | 74.7 | - | 0.204 | - |
| | – MFD [21] | 74.7 | - | 0.178 | - |
| | **Self-Supervised** | | | | |
| | – SimCLR [6] | 78.0 $\pm$ 0.25 | 0.335 $\pm$ 0.03 | 0.421 $\pm$ 0.04 | 0.287 $\pm$ 0.03 |
| | – **CCL (Ours)** | **78.5** $\pm$ 0.22 | **0.191** $\pm$ 0.02 | **0.156** $\pm$ 0.02 | **0.089** $\pm$ 0.01 |
| | – **Tight-CCL (Ours)** | **78.3** $\pm$ 0.15 | **0.188** $\pm$ 0.02 | **0.159** $\pm$ 0.02 | **0.110** $\pm$ 0.02 |

### B.9 Limitations, discussions and future work

One important future work direction, and a limitation of this work, is to study the scenario the sensitive information is unknown or partially known. It could be addressed by using other auxiliary attributes (e.g., image annotations or captions) in the datasets that are highly relevant to the sensitive attributes, or first train a separate model to capture bias features and then train the main model by learning features orthogonal to the bias feature. Another important problem is to remove the effect of multiple sensitive attributes simultaneously, which may be addressed by using a joint distribution of multiple sensitive attributes. If there are too many sensitive attributes, we can perform a dimensional reduction.

For the social impact, CCL may bring a positive impact by removing gender, race, or identity information from representations. The potential negative impact is that this method could be intentionally used to remove information that should be available and included in the representation, for example, gender information in a model for medical diagnosis.