Reducing Representation Bias through Fairness-Driven Sampling in Contrastive Learning

David Martin[†,*], Blessing Ogbuokiri[‡]
[†] Responsible and Applied Machine Learning Laboratory (RAML Lab),
Department of Computer Science, Brock University

Abstract

Contrastive learning is a widely applicable Self-Supervised machine learning algorithm that has demonstrated state of the art performance often competing with supervised learning methods. However, the stochastic approach to sampling can inherently amplify representation bias, as over-represented groups are more likely to dominate contrastive pair construction while underrepresented groups receive limited exposure during training leading to imbalanced subgroup representation and biased downstream performance. To address this issue, we propose a fairness-driven sampling algorithm that leverages latent similarity structure to infer subgroup information and guide positive and negative pair selection without the reliance on annotated demographic attributes. Our fairness-driven approach is evaluated in terms of both fairness representation and utility. The results show that our fairness-driven sampling strategy not only increases representation across underrepresented latent subgroups, but maintains competitive accuracy with baseline Contrastive learning sampling. This method has the potential to improve fairness in downstream applications such as facial recognition, clinical diagnostics, and language models deployed in demographically diverse or low-resource contexts.

Keywords: Contrastive Learning, SimCLR, Representation Bias, Bias mitigation, Responsible AI

1. Introduction

Contrastive learning has proven to be an efficient approach in self-supervised representation learning, with growing applications in domains such as medical imaging [1] and facial recognition [2]. Contrastive learning frameworks including SimCLR and MoCo have demonstrated performance comparable to supervised learning methods [3, 4]. However, recent studies have demonstrated that contrastive learning can incorporate sensitive demographic information resulting in biased downstream performance within high-stakes domains including healthcare [5] and facial recognition [6].

These biases can be aresult of traditional sampling strategies in which positive and negative pairs are sampled stochastically, often leading to imbalanced subgroup representations and biased downstream performance, particularly in scenarios where demographic labels are not available. Although existing fairness-aware sampling strategies have demonstrated improvements in fairness outcomes in contrastive learning, previous approaches rely on explicit demographic sensitive attributes limiting their applicability to unlabeled environments [7].

Our proposed sampling method aims to balance the representation of inferred subgroups in the absence of demographic attributes. By leveraging latent subgroup discovery from the initial contrastive embeddings via clustering, we define a balanced sampling distribution across latent subgroups. For each anchor sample, positive pairs are formed using Euclidean distance to identify the closest embedding of the same subgroup. Negative pairs are constructed by sampling from different subgroups, prioritizing underrepresented clusters, and selecting samples with significant distance in the representation space to preserve strong contrast between dissimilar pairs.

dm20zo@brocku.ca

We evaluated our sampling method on two benchmark datasets, CelebFaces attributes (CelebA [8]) containing 202,599 celebrity images with 40 annotated attributes, and CheXpert-v1.0-small (CheXpert[9]) containing 224,316 chest radiographs. Compared with a random sampling baseline, our fairness-driven sampling method demonstrates improvements in fairness metrics with minimal utility degradation.

2. Related Work

2.1. NT-Xent Loss

The normalized temperature-scaled cross entropy loss (NT-Xent) is the standard objective function for the SimCLR framework [3]. The contrastive framework depends on this objective function by drawing similar (positive) pairs together and pushing dissimilar (negative) pairs apart. This is illustrated in equation 2.1

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{2.1}$$

where $\mathbf{z}_i$ represents the anchor embedding with its corresponding positive embedding $\mathbf{z}_j$, $\mathbf{z}_k$ represents all other embeddings in the batch serving as negative samples, and τ represents a temperature parameter. In standard SimCLR implementations, negative samples are selected uniformly from the batch. However, our fairness-driven sampling approach explicitly applies weights to negative samples according to their cluster membership, ensuring proportional exposure across latent subgroups during training.

2.2. Prior Work: Fairness in Contrastive Learning

Previous work has addressed such fairness gaps in self-supervised contrastive learning. Conditional Contrastive Learning (CCL)[7] introduces a sampling heuristic that conditions positive and negative pair selection on sensitive demographic attributes, by strictly forming pairs within the same gender or race. This reduces the influence of demographic attributes as the model no longer learns to distinguish between different demographic groups. DualFair [10] extends this by optimizing for both group-level and individual-level fairness, selecting negative samples from the same demographic group while treating a sample and its attribute-flipped counterpart as positive pairs. While both sampling frameworks significantly improve fairness of contrastive self-supervised models, both require access to sensitive labels during training. Our sampling method addresses this limitation by discovering latent subgroups through performing unsupervised clustering on the generated representation space, requiring no demographic annotations.

3. Methodology

We present a fairness-driven sampling strategy method that mitigates representation bias without requiring access to demographic labels during training. In this section, we elaborate on each component of the sampling algorithm and the hyperparameters selected for our experiments. Hardware details for each experiment is shown in Table 5, Appendix A.4.

3.1. Sampling Algorithm Design

Our fairness-driven sampling algorithm is comprised of two separate phases: latent subgroup discovery and fairness-driven pair construction.

(1) Phase 1 - Latent Subgroup discovery: Without access to demographic labels, we first trained a SimCLR framework with a ResNet-50 backbone encoder followed by

a two-layer projection head with a ReLU activation function projecting to a 128-dimensional embedding space. For each anchor image, two correlated views are generated through a set of stochastic data augmentations including, color jittering, random cropping, and grayscale conversions complying with best practices in contrastive learning frameworks [3, 4]. The model was trained with the temperature-scaled NT-Xent loss with a temperature of $\tau = 0.5$ and a batch size of 256 over 30 epochs. To identify latent subgroups without demographic labels, we applied K-means clustering with $K = 4$ to the baseline embeddings. Each sample was then assigned a cluster label $c_i \in \{0, 1, 2, 3\}$ based on cluster membership.

(2) Phase 2 - Fairness-driven pair construction: Our fairness-driven sampling algorithm constructs contrastive pairs by first quantifying subgroup representation in the embedding space. For each cluster c we calculate $p(c) = N_c/N$, where N_c is the number of samples assigned to cluster c and N is the total number of samples. We then establish the uniform target distribution$q(c) = 1/C$ where C is the total number of clusters. The difference between $p(c)$ and $q(c)$ measures the subgroup representation gap. Clusters where $p(c) < q(c)$ are underrepresented, while clusters where $p(c) > q(c)$ are overrepresented and will be downweighted. Our sampling strategy for pair construction is as follows:

- Positive Pairs: For each anchor embedding z_i with cluster assignment c_i, we construct a positive pair by selecting its nearest neighbor within the same cluster using the Euclidean distance formula $||z_i - z_j||_2$ between that anchor and all candidate embeddings in the same cluster. See Algorithm 1 (lines 5–7), Appendix A.1.
- Negative Pairs: For negative pair construction, we first assign weights to all clusters of which the anchor is not a member, based on their representation gap $w(c) = \max(0, q(c) - p(c)) + \epsilon$, where underrepresented clusters ($p(c) < q(c)$) receive higher weights proportional to their deficit, while overrepresented clusters receive only the baseline weight ϵ where $\epsilon = 0.1$. These weights are then normalized to form probabilities. We then apply distance based sampling favoring embeddings that are farther from the anchor, ensuring that underrepresented latent subgroups receive proportional exposure during training. See Algorithm 1 (lines 8–16), Appendix A.1.

Our sampling approach essentially aims to address bias in standard contrastive learning by ensuring that all latent subgroups receive equal sample exposure during training, preventing majority clusters from dominating pair construction at the expense of underrepresented subgroups.

4. **Results**

We evaluate our fairness-driven contrastive learning approach on CelebA for facial attribute classification and CheXpert for medical imaging, assessing performance in terms of both fairness improvements in learned representations and downstream task performance.

4.1. **Baseline Embedding Analysis**

The baseline SimCLR embeddings reveal clustering patterns across both datasets, with CelebA forming more balanced clusters compared to CheXpert. See Figure 1 and Table 3, Appendix A.2. CheXpert exhibits more severe cluster imbalance, with Cluster 1 dominating at 43.9%, illustrating the need for fairness-driven sampling to ensure exposure of underrepresented groups during training.

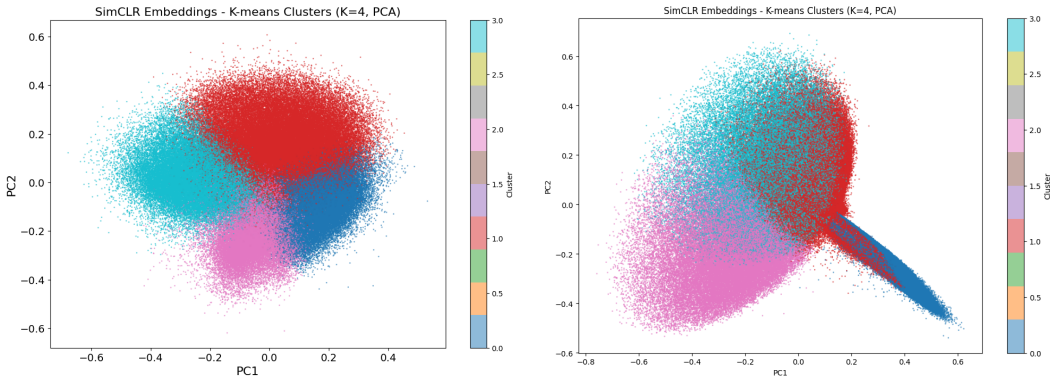

Figure 1. Baseline SimCLR embeddings using PCA for CelebA (left) and CheXpert (right).

4.2. **Utility & Fairness Metrics**

This section is a comparative analysis on fairness and utility metrics between the baseline sampler and our fairness-driven sampling strategy implemented into the SimCLR framework.

4.2.1. **Fairness Metrics**

We measure fairness using three metrics that capture aspects of cluster balance and quality. Cluster dominance ratio, which quantifies the number of samples in the largest cluster where lower values indicate more balanced cluster sizes. Representation entropy measures the uniformity of the cluster size distribution. Subgroup separation distance which computes the average Euclidean distance between centroids, where higher values indicate more distinct subgroup representations. Results are shown in Table 1 and Figure 2.

Dataset	**Fairness Metric**	**Baseline**	**Fairness-Driven**	**Change (%)**
CelebA	Cluster Dominance Ratio ↓	0.3569	0.3398	-4.79%
	Representation Entropy ↓	1.9463	1.9070	-2.02%
	Subgroup Separation Distance ↑	0.4331	0.6212	+43.43%
CheXpert	Cluster Dominance Ratio ↓	0.4394	0.3943	-10.26%
	Representation Entropy ↓	1.8314	1.8092	-1.21%
	Subgroup Separation Distance ↑	0.5922	0.9324	+57.45%

Table 1. Fairness Metrics Comparison: Baseline vs Fairness-Driven SimCLR for CelebA and CheXpert.

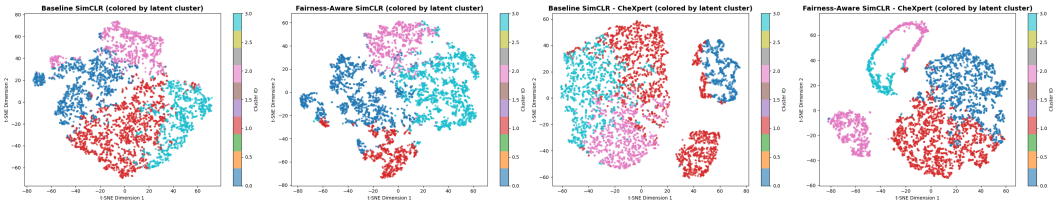

Figure 2. Baseline sampler SimCLR vs Fairness sampler SimCLR for CelebA(Left) and CheXpert(Right). Both T-SNE visualizations show significant improvement in cluster separation.

Our fairness-driven sampling strategy demonstrates consistent improvements across both datasets, with a decrease in cluster dominance ratio indicating more balanced cluster sizes,

and a dramatic increase in subgroup separation distance indicating greater distinction between underrepresented subgroups. Though representation entropy decreased slightly in both datasets, it was roughly only by 1-2 %, it remains near the theoretical maximum, representing a minor trade off for substantial gains in cluster separation. This is visually apparent in Figure 2.

4.2.2. **Utility Metrics**

To assess utility, we conducted linear probe evaluation by training logistic regression classifiers on the embeddings from both the baseline sampler and our fairness-driven sampler following standard practice in self-supervised learning [3, 4]. For CelebA, we evaluated performance on two attributes: Attractive and Wavy hair. For Chexpert we tested on two pathology detection tasks: Cardiomegaly and Edema. Both classifiers had a learning rate of 0.0001, and a simple 80-20 train/test split for fair comparison. This is illustrated in Table 2

Dataset	**Attribute**	**Metric**	**Baseline**	**Fairness-Driven**	**Δ**
CelebA	Attractive	Accuracy	0.7687	0.7607	-0.0080
		F1-Score	0.7824	0.7763	-0.0061
		AUC-ROC	0.8456	0.8360	-0.0096
	Wavy Hair	Accuracy	0.7325	0.7233	-0.0092
		F1-Score	0.5365	0.4915	-0.0450
		AUC-ROC	0.7838	0.7694	-0.0144
CheXpert	Cardiomegaly	Accuracy	0.8790	0.8790	0.0000
		F1-Score	0.0011	0.0000	-0.0011
		AUC-ROC	0.7137	0.6972	-0.0165
	Edema	Accuracy	0.7799	0.7791	-0.0009
		F1-Score	0.3417	0.3343	-0.0075
		AUC-ROC	0.7844	0.7796	-0.0048

Table 2. Downstream Classification Performance: Baseline Sampler vs Fairness-Driven Sampler for CelebA and CheXpert. Linear probes trained on embeddings.

The downstream classification results illustrate that the integration of our fairness driven sampling algorithm had minimal impact on model utility. For CelebA, average accuracy decreased by less than 1%. CheXpert exhibited higher utility preservation with Cardiomegaly showing no accuracy alterations and Edema experiencing only a 0.0009 in accuracy loss followed by minimal decrease in F1-score and AUC-ROC for both attributes, see Table 2. This is consistent across three other pathologies in Table 4, Appendix A.3. The most notable degradation occurs in the F1-score and AUC-ROC for CelebA's wavy hair likely due to having a lower baseline performance. However, these results remain competitive suggesting our fairness-driven sampling strategy reduces representation bias without compromising on utility of learned embeddings for downstream tasks.

5. **Conclusion**

In this paper, we presented a fairness-driven sampling strategy. We show that it can be implemented into notable contrastive learning frameworks such as SimCLR and MoCo while demonstrating a substantial increase in underrepresented subgroup representation in addition to minimal compromise to utility across two distinct datasets on multiple attributes. We hope this work opens new perspectives on bias mitigation in contrastive learning with practical applications in domains such as facial recognition and medical imaging diagnosis where demographic equity is essential.

Acknowledgment

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2026-05435 and DGECR-2026-00353]. The authors also acknowledge the support of the Responsible and Applied Machine Learning Lab (RAML Lab) at Brock University's Department of Computer Science for providing a collaborative research environment and resources that contributed to this work.

Declaration of AI Use

The authors used ChatGPT to improve language and readability. All content was subsequently reviewed and edited, and the authors take full responsibility. No AI tools were used to generate scientific content, data, analysis, or conclusions.

References

[1] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi. *Big Self-Supervised Models Advance Medical Image Classification*. 2021. arXiv: 2101.05224 [eess.IV]. URL: https://arxiv.org/abs/2101.05224.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8748–8763.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*. 2020. arXiv: 2002.05709 [cs.LG]. URL: https://arxiv.org/abs/2002.05709.

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[5] L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi. "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations". In: *Nature medicine* 27.12 (2021), pp. 2176–2182.

[6] J.-R. Conti, N. Noiry, S. Clemencon, V. Despiegel, and S. Gentric. "Mitigating gender bias in face recognition using the von mises-fisher mixture model". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 4344–4369.

[7] M. Q. Ma, Y.-H. H. Tsai, P. P. Liang, H. Zhao, K. Zhang, R. Salakhutdinov, and L.-P. Morency. *Conditional Contrastive Learning for Improving Fairness in Self-Supervised Learning*. 2022. arXiv: 2106.02866 [cs.LG]. URL: https://arxiv.org/abs/2106.02866.

[8] Z. Liu, P. Luo, X. Wang, and X. Tang. *Deep Learning Face Attributes in the Wild*. 2015. arXiv: 1411.7766 [cs.CV]. URL: https://arxiv.org/abs/1411.7766.

[9] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[10] S. Han, S. Lee, F. Wu, S. Kim, C. Wu, X. Wang, X. Xie, and M. Cha. "DualFair: Fair representation learning at both group and individual levels via contrastive self-supervision". In: *Proceedings of the ACM web conference 2023*. 2023, pp. 3766–3774.

Appendix A.

A.1. Fairness-aware Sampling Algorithm

Algorithm 1 Fair Contrastive Pair Sampling

Require: Embeddings $\mathbf{Z}$, cluster labels $\mathbf{C}$, fairness parameter ε
Ensure: Positive pair set P, negative pair set N
Initialize $P \leftarrow \emptyset$, $N \leftarrow \emptyset$
Compute $p(c) = N_c/N$ for each cluster c; set $q(c) = 1/C$
for each anchor embedding $z_i \in \mathbf{Z}$ **do**
  Select cluster assignment $c_i = C[i]$
  Positive Pair Sampling
  Select z_j from cluster c_i such that $\|z_i - z_j\|_2$ is minimized
  Add (z_i, z_j) to P
  Negative Pair Sampling
  for each cluster $c \neq c_i$ **do**
    Compute sampling weight: $w(c) = \max\big(0,\ q(c) - p(c)\big) + \varepsilon$
  end for
  Normalize $w(c)$ to form probability distribution
  Sample z_k proportional to $w(c)$ and $\|z_i - z_k\|_2$
  Add (z_i, z_k) to N
end for
return P, N

A.2. Cluster Distributions for CelebA and CheXpert

CelebA			**CheXpert**		
Subgroup	**Count**	**%**	**Subgroup**	**Count**	**%**
Cluster Distribution					
Cluster 0	53,653	26.5%	Cluster 0	23,371	10.4%
Cluster 1	72,311	35.7%	Cluster 1	98,267	43.9%
Cluster 2	36,428	18.0%	Cluster 2	46,026	20.6%
Cluster 3	40,207	19.8%	Cluster 3	55,984	25.0%
Total	**202,599**	**100%**	**Total**	**223,648**	**100%**

Table 3. Baseline Cluster Distributions for CelebA and CheXpert

A.3. Baseline sampler vs Fairness-Driven Sampler Utility Metrics on CheXpert pathologies

Attribute	Metric	Baseline	Fairness-Driven	Δ
Consolidation	Accuracy	0.9338	0.9338	0.0000
	F1-Score	0.0000	0.0000	0.0000
	AUC-ROC	0.6618	0.6504	-0.0114
Atelectasis	Accuracy	0.8504	0.8504	0.0000
	F1-Score	0.0000	0.0000	0.0000
	AUC-ROC	0.6153	0.6092	-0.0061
Pleural Effusion	Accuracy	0.6882	0.6757	-0.0125
	F1-Score	0.5571	0.5565	-0.0006
	AUC-ROC	0.7402	0.7235	-0.0167

Table 4. Downstream classification performance for CheXpert: baseline vs fairness-driven sampler across five pathology detection tasks withi minimal degradation.

A.4. Experiment Hardware Setup

Component	Specification
CPU	Intel Xeon @ 2.00GHz (4C/8T)
GPU	NVIDIA A100-SXM4 (80 GB VRAM)
CUDA	12.4
Environment	Google Colab (KVM)

Table 5. System Specifications.