REFERENCES

Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. *Advances in neural information processing systems*, 28, 2015.

Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. *arXiv preprint arXiv:2211.05568*, 2022.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111, 2020.

Junyi Chai and Xiaoqian Wang. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems*, 35:27100–27113, 2022.

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

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders. *arXiv preprint arXiv:2103.06413*, 2021.

R Child, S Gray, A Radford, and I Sutskever. Generating long sequences with sparse transformers. arxiv 2019. *arXiv preprint arXiv:1904.10509*.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://www.aclweb.org/anthology/W19-4828.

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*, pp. 1436–1445. PMLR, 2019.

Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pp. 3662–3666. IEEE, 2020.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

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*, pp. 770–778, 2016.

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*, pp. 9729–9738, 2020.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://www.aclweb.org/anthology/D19-1275.

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.

Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34:26449–26461, 2021.

Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10357, 2022.

Ahmad Khajehnejad, Moein Khajehnejad, Mahmoudreza Babaei, Krishna P Gummadi, Adrian Weller, and Baharan Mirzasoleiman. Crosswalk: Fairness-enhanced node representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 11963–11970, 2022.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frederic Dreyer, Aleksandar Shtedritski, and Yuki Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. *Advances in neural information processing systems*, 34:2611–2624, 2021.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Hongyi Ling, Zhimeng Jiang, Youzhi Luo, Shuiwang Ji, and Na Zou. Learning fair graph representations via automated data augmentations. In *The Eleventh International Conference on Learning Representations*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Zheqi Lv, Wenqiao Zhang, Shengyu Zhang, Kun Kuang, Feng Wang, Yongwei Wang, Zhengyu Chen, Tao Shen, Hongxia Yang, Beng Chin Ooi, et al. Duet: A tuning-free device-cloud collaborative parameters generation framework for efficient device model generalization. In *Proceedings of the ACM Web Conference 2023*, pp. 3077–3085, 2023.

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

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

Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10389–10398, 2022.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 189–198. IEEE, 2018.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The annals of mathematical statistics*, pp. 832–837, 1956.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Contrastive learning for fair representations. *arXiv preprint arXiv:2109.10645*, 2021.

Anshumali Shrivastava and Ping Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). *Advances in neural information processing systems*, 27, 2014.

Kyungwoo Song, Yohan Jung, Dongjun Kim, and Il-Chul Moon. Implicit kernel attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9713–9721, 2021.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Qinghua Tao, Francesco Tonin, Panagiotis Patrinos, and Johan AK Suykens. Nonlinear svd with asymmetric kernels: feature learning and asymmetric nystr\" om method. *arXiv preprint arXiv:2306.07040*, 2023.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://www.aclweb.org/anthology/P19-1452.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: a unified understanding of transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Han Zhao, Louis-Philippe Morency, and Ruslan Salakhutdinov. Self-supervised representation learning with relative predictive coding. *arXiv preprint arXiv:2103.11275*, 2021a.

Yao-Hung Hubert Tsai, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning: Removing undesirable information in self-supervised representations. *arXiv e-prints*, pp. arXiv–2106, 2021b.

Yao-Hung Hubert Tsai, Tianqin Li, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. Conditional contrastive learning with kernel. *arXiv preprint arXiv:2202.05458*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 63–76, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4808. URL https://www.aclweb.org/anthology/W19-4808.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://www.aclweb.org/anthology/P19-1580.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5310–5319, 2019.

Matthew A Wright and Joseph E Gonzalez. Transformers are deep infinite-dimensional non-mercer binary kernel machines. *arXiv preprint arXiv:2106.01506*, 2021.

Mike Wu, Milan Mosse, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. Conditional negative sampling for contrastive learning of visual representations. *arXiv preprint arXiv:2010.02037*, 2020.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

Cindy Zhang, Sarah Huiyi Cen, and Devavrat Shah. Matrix estimation for individual fairness. In *International Conference on Machine Learning*, pp. 40871–40887. PMLR, 2023.

Fengda Zhang, Kun Kuang, Long Chen, Yuxuan Liu, Chao Wu, and Jun Xiao. Fairness-aware contrastive learning with partially annotated sensitive attributes. In *The Eleventh International Conference on Learning Representations*, 2022.

Appendix for "Fairness-Aware Attention for Contrastive Learning"

Table of Contents

A Experimental Details

This section provides the details of the model and training for experiments in Section 4.

A.1 Training and Evaluation

ColorMNIST. Samples in the colorMNIST dataset are 32x32 resolution handwritten digit images, where the digit is represented in black and the background is some known assigned color which is representable as a continuous RGB color vector. The train-test split is 60,000 training images to 10,000 test images. The augmentation scheme is randomized resized crop followed by a random horizontal flip. We pre-train using the LARS optimizer (You et al. (2017)) and cosine annealing for the learning rate scheduler. The full FARE attention mechanism with sparsification uses 8 rounds of hashing, a bucket size of 64, and backwards and forwards cross-bucket attention. The linear classifier is trained using L-BFGS as optimizer over 500 iterations. We pre-train with a batch size of 256 for 50 epochs.

Figure 2: colorMNIST dataset (Tsai et al., 2022)

We follow the recent contrastive learning literature (Chen et al. (2020), Robinson et al. (2020), Wu et al. (2020)) and pre-train the full model before discarding everything except the backbone encoder at evaluation time.

CelebA. The train-test split is the default as provided by PyTorch. Images are resized to 128×128. Resnet-18 (He et al., 2016) is the encoder and we use the same 2-layer MLP and random augmentation strategies as Chen et al. (2020). Same as with colorMNIST, we pre-train with the LARS optimizer and use cosine annealing. We use a batch size of 512 and the LSH scheme uses buckets of size 128 with 8 rounds of hashing and backwards and forwards cross-bucket attention. We train the full model for 100 epochs and evaluate with a single linear layer trained on the frozen encodings for 10 epochs using Adam as optimizer.

To evaluate the fairness of the representations, we adopt the Equalized Odds (EO) metric (Hardt et al., 2016). Following Jung et al. (2022) and Zhang et al. (2022), we compute the metric over multiple sensitive attributes by:

$$\max_{\forall s^i, s^j \in S} \overline{\sum}_{\forall y, \hat{y}} \left| P_{s^i}\left(\hat{Y} = \hat{y} | Y = y\right) - P_{s^j}\left(\hat{Y} = \hat{y} | Y = y\right) \right|, \tag{14}$$

where $\overline{\sum}$ is the averaged sum, Y is the target label, $\hat{Y}$ is the predicted label, and $s_i, s_j \in S$ are values of sensitive attributes. A smaller EO means a fairer model.

A.2 Baselines

ColorMNIST. The relevant baselines for comparison are the InfoNCE model (**InfoNCE**) (Oord et al., 2018), the Fair-InfoNCE model with clustering (**Fair-InfoNCE**) (Tsai et al., 2022) and the conditional contrastive learning with kernel model (**CCLK**) (Tsai et al. (2022)).

The InfoNCE model uses the InfoNCE loss function 15 without performing any conditional sampling. The Fair-InfoNCE model uses the Fair-InfoNCE loss function 21 and performs conditional sampling by first clustering the protected attribute so as to discretize it and then sampling from within the same cluster as the anchor. We report this model's results according to its best performing cluster size as determined by its authors, which is found to be a 10-cluster partition. CCLK uses a kernel similarity metric for weighing negative samples in the batch according to their similarity in the bias-dimension. We report its results according to its best performing kernel choice as chosen by its authors which was the cosine kernel.

The InfoNCE objective (Oord et al., 2018) used in the baseline model InfoNCE is given by:

$$\sup_{f} \mathbb{E}_{(x, y_{pos}) \sim P_{XY},\ \{y_{neg}\}_{i=1}^{n} \sim P_Y^{\otimes n}} \left[\log \frac{e^{f(x, y_{pos})}}{e^{f(x, y_{pos})} + \sum_{i=1}^{b} e^{f(x, y_{neg,i})}} \right] \tag{15}$$

CelebA. We compare with **SimCLR** (Chen et al., 2020) and all kernel implementations of CCLK provided by Tsai et al. (2022). For each kernel model, the kernel in the name refers to the what kernel similarity metric is chosen for measuring the similarity across protected attributes, which then determines the relevance of that sample for being contrasted with the positive sample. For example, CCLK-RBF uses the RBF kernel to compute similarity between two protected attributes.

B Connection between Kernel-based Scoring Function Estimation in (Tsai et al., 2022) and Attention

The CCLK model uses the following kernel-based scoring function estimation:

Proposition 2 (Kernel-Based Scoring Function Estimation (Tsai et al., 2022))**.** *Given* $\{x_i, y_i, z_i\}_{i=1}^{b} \sim P_{XYZ}^{b}$*, the similarity score of the data pair* (x_i, y) *given the anchor* z_i *is computed via the finite-sample kernel estimation* $e^{f(x_i, y)}$ *when* $y \sim P_{Y|Z=z_i}$ *as follows:*

$$e^{f(x_i, y)} = \left[K_{XY} (K_Z + \lambda \boldsymbol{I})^{-1} K_Z \right]_{ii}, \tag{16}$$

for $i = 1, \ldots, b$, $[K_{XY}]_{ij} := e^{f(x_i, y_j)}$, and $[K_Z]_{ij} := \langle \gamma(z_i), \gamma(z_j) \rangle_{\mathcal{G}}$, where γ is some kernel feature embedding, $\mathcal{G}$ is the corresponding Reproducing Kernel Hilbert Space (RKHS), and $\langle \cdot, \cdot \rangle_{\mathcal{G}}$ is an inner product in space $\mathcal{G}$.

First, in comparison to Eqn. 2, FARE and sparseFARE avoid matrix inversion. FARE's attention computation has complexity $O(b^2)$ (Vaswani et al., 2017) and sparseFARE has complexity $O(b\ \ logb)$ (Kitaev et al., 2020), which improve significantly over $O(b^3)$ in Eqn. 2.

Second, our methods do not impose assumptions on the bias-causing interactions over protected attributes. In particular, we avoid specifying any particular kernel and allow our attention mechanism to learn the bias-causing interactions. To see this difference, we decompose the estimator in Eqn. 2 as follows:

$$\begin{aligned} &\left[K_{XY}(K_Z + \lambda \mathbf{I})^{-1} K_Z\right]_{ii} \qquad (17)\\ &= [K_{XY}]_{i*}[(K_Z + \lambda \mathbf{I})^{-1} K_Z]_{*i} \\ &= \sum_{j}^{b} w(z_i, z_j) e^{f(x_i, y_j)}, \qquad (18) \end{aligned}$$

where $w(z_i, z_j) = [(K_Z + \lambda I)^{-1} K_Z]_{ij}$ are smoothed kernel similarity scores (Tsai et al., 2022). Hence we see the (Tsai et al., 2022) estimator as performing a similar weighting of similarity scores between samples, with weights provided by the similarities over the protected attributes. This approach differs from ours however since the kernel must be pre-specified in K_Z. This imposes strong assumptions on bias-causing interactions that limit the extent to which the model can learn fair representations. Our method by contrast can be understood as replacing $w(z_i, z_j)$ with attention score $p(z_i, z_j)$. The attention mechanism can more flexibly model the bias-causing interactions and learns to focus-attention on bias-reducing samples that help learn the representation space.

We provide a proof adapted from (Tsai et al., 2022) of their kernel-based scoring function estimation below.

Proof of kernel-based scoring function estimation. First, letting $\Phi = [\phi(g(y_1)), \ldots \phi(g(y_b))]^\top$ be the matrix of kernel embeddings for encodings $g(y_i)$ with feature map ϕ and $\Gamma = [\gamma(z_1), \ldots, \gamma(z_b)]^\top$ be the matrix of kernel embeddings for protected attribute outcomes z with feature map γ, Definition 3 provides the Kernel Conditional Embedding Operator (Song et al., 2013):

Definition 3. *[Kernel Conditional Embedding Operator (Song et al., 2013)] The finite-sample kernel estimation of $\mathbb{E}_{y \sim P_{Y|Z=z}}\left[\phi(g(y))\right]$ is $\Phi^\top (K_Z + \lambda \boldsymbol{I})^{-1} \Gamma \gamma(z)$ where λ is a hyperparameter.*

Then, according to Definition 3, for any given $Z = z$, $\phi(g(y))$ when $y \sim P_{Y|Z=z}$ can be estimated by

$$\Phi^\top (K_Z + \lambda \mathbf{I})^{-1} \Gamma \gamma(z) \qquad (19)$$

We look for the inner product between (5) and the encoding of (x_i, z_i) when $y \sim P_{Y|Z=z_i}$:

$$\begin{aligned} &\langle \phi(g(x_i)), \Phi^\top (K_Z + \lambda \mathbf{I})^{-1} \Gamma \gamma(z_i) \rangle_{\mathcal{H}} = tr\Big(\phi(g(x_i)\Big)^\top \Phi^\top (K_Z + \lambda \mathbf{I})^{-1} \Gamma \gamma(z_i) \\ &= [K_{XY}]_{i*} (K_Z + \lambda \mathbf{I})^{-1} [K_Z]_{i*} = [K_{XY}]_{i*} [(K_Z + \lambda \mathbf{I})^{-1} K_Z]_{*i} \\ &= [K_{XY}(K_Z + \lambda \mathbf{I})^{-1} K_Z]_{ii} \qquad (20) \end{aligned}$$

□

C Comparison of Fair-InfoNCE and FAREContrast

We present a discussion of the differences between the Fair-InfoNCE objective from Tsai et al. (2021b) and the FAREContrast objective we use to train our attention-based FARE models. FAREContrast is derived from Fair-InfoNCE by replacing the conditionally sampled negative pairs with the output of the FARE attention mechanism. This leads to a difference firstly in sampling procedure and secondly in the inclusion of learnable attention scores in the loss.

The Fair-InfoNCE (Tsai et al., 2021b) is given as:

$$\sup_f \mathbb{E}_{z \sim P_Z,\ (x, y_{pos}) \sim P_{XY|Z=z},\ \{y_{neg}\}_{i=1}^b \sim P_{Y|Z=z}^{\otimes b}} \left[\log \frac{e^{f(x, y_{pos})}}{e^{f(x, y_{pos})} + \sum_{i=1}^b e^{f(x, y_{neg,i})}}\right], \qquad (21)$$

and FAREContrast is given as:

$$\sup_f \mathbb{E}_{\{(x_i,y_i,z_i)\}_{i=1}^b \sim P_{XYZ}^{\otimes b}} \left[\log \frac{e^{f(x_i,y_i)}}{e^{f(x_i,y_i)} + \sum_{j=1}^b \text{softmax}\left((W_Q z_i)^\top W_K z_j/\rho\right) e^{f(x_i,y_j)}} \right], \quad (22)$$

where b denotes the batch size, $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a mapping given by $f(x,y) = \text{cosine similarity}\Big(g_{\theta_X}(x), g_{\theta_Y}(y)\Big)/\tau$, $g_{\theta_X}, g_{\theta_Y}$ are neural networks parameterized by θ_X, θ_Y, and τ is a hyperparameter scaling the cosine similarity.

We see that FAREContrast does not require conditional sampling of the negatively paired samples, $\{y_{neg}\}_{i=1}^b \sim P_{Y|Z=z}^{\otimes b}$ for outcome of the of the protected attribute z. Instead, FARE considers the whole batch and selectively weights samples according to their protected attribute status. One issue with conditional sampling as in Eqn. 21 is data scarcity, whereby conditioning on $Z = z$ can lead to insufficient negative samples for contrasting (Tsai et al., 2022). This problem is exacerbated when the protected attribute has high cardinality or is continuous, which is the problem setting we aim to deal with. When there are insufficient negative samples, we incur risk of poorly learnt representations and collapse (Chen et al., 2020; Chen & He, 2021). For this reason, we derive FARE which considers the whole batch and uses learnt attention scores to accentuate/attenuate negative samples according to their bias characteristics.

The second difference is then the attention weights included in FAREContrast. Including the attention weights in FAREContrast means that that FARE learns according to information coming from the gradients and so can better focus on samples that help minimize the loss, thereby helping the encoder to learn meaningful representations.

D Additional Results

D.1 LSH Bucket Scheme

Attention Scheme	Top-1 Test Accuracy (↑)	Bias Removal (↑)
Adjacent	86.4 ± 1.3	74.0 ± 3.8
Intra	84.9 ± 2.1	58.2 ± 9.8

Table 3: Sparsification Scheme on ColorMNIST Results. Bias removal is measured by MSE, where high MSE indicates more color information has been removed from the learned representations.

Table 3 shows results for when the LSH scheme considers intra-bucket attention versus the standard adjacent bucket attention (where attention is computed across adjacent buckets). We see fairly substantial drop in performance when restricting attention to within the same bucket, both in terms of accuracy and fairness. Lower accuracy is intuitive given the intra-bucket attention removes three quarters of negative samples, which depletes the model's ability to learn meaningful representations. At the same time, we see lower fairness, despite the heavy debiasing scheme. This may support the conclusion that to learn effectively debiased representations, the model needs sufficiently many samples to learn to attend over and focus on bias-reducing samples. With too few samples in the batch, the model is ignoring too many samples, including ones that would help it learn debiased representations.

D.2 Fairness-Accuracy Tradeoff

The two metrics that capture both representation quality and fairness are Accuracy and Equalized Odds (EO). Table 2 showed that SparseFARE Pareto dominates all kernel baselines in terms of both fairness and accuracy, with the exception of CCLK-Linear and CCLK-Polynomial, which were able to attain slightly higher accuracy. We therefore further compare SparseFARE to these two models by plotting the fairness-accuracy tradeoff curves in Figure **??**. The curves are produced by plotting EO and Accuracy at four stages during training - after 25, 50, 75, and 100 epochs. We see that for every level of accuracy, SparseFARE achieves better fairness (lower EO). This implies that SparseFARE attains a better fairness-accuracy tradeoff. Additionally of interest, we find that

SparseFARE is even able to simultaneously minimize EO while increasing accuracy, implying that it can learn representations that do not necessarily need to compromise fairness for higher accuracy.

D.3 Comparison With Work in Partial Access to Sensitive Attributes

Model	Test Accuracy (↑)	EO (↓)
Supervised Models		
CGL + G-DRO (Sagawa et al., 2019)	71.4	21.9
CGL + FSCL (Park et al., 2022)	74.0	25.6
Unsupervised Models		
CGL + VFAE (Louizos et al., 2015)	72.7	28.7
CGL + GRL (Raff & Sylvester, 2018)	73.8	26.9
SimCLR (Chen et al., 2020)	**77.7**	39.6
FairCL (Zhang et al., 2022)	74.1	24.5
FARE (**ours**)	73.7	23.5
SparseFARE (**ours**)	70.4	**18.7**

Table 4: CelebA Results. Fare and SparseFARE in comparison with unsupervised and supervised models under partial sensitive label access.

This paper uses the same experimental setup on CelebA as Zhang et al. (2022) in terms of training procedure and evaluation protocol. Zhang et al. (2022) differs, however, in the sense that the authors assumes only partial access to sensitive attributes and therefore use auxiliary models, for example an editor (Zhang et al., 2022) or CGL (Jung et al., 2022), to solve this problem. Given the experimental setups are the same, we include their results as well for reference, however we do not feature these results in the main body given the important difference regarding sensitive attribute access.

E Fair Attention-Contrastive Criterion

We do not include a learnable value transformation W_V on the raw similarity scores such that $V = UW_V$ where $U = [e^{f(x_i,y_j)}]_{ij}$ as doing so allows the optimization process to obtain 0 loss without learning meaningful representations. This is seen immediately from the criterion, where allowing W_V gives individual similarity scores as $w_{ij}e^{f(x_i,y_j)}$ in the criterion:

$$\sup_f \mathbb{E}_{\{(x_i,y_i,z_i)\}_{i=1}^b \sim P_{XYZ}^b} \left[\log \frac{e^{f(x_i,y_i)}}{e^{f(x_i,y_i)} + \sum_{j \in S_i} p(z_i,z_j) w_{ij} e^{f(x_i,y_j)}} \right]$$

hence the loss is minimised by sending $w_{ij} \to \infty \ \ \forall i,j$.

F ETHICAL CONSIDERATIONS

We note that there are two, interconnected prevalent ethical issues in fair ML. The first is that almost all fair ML literature simplifies the problem of fairness to simple binaries and the second is that fairness metrics (which are typically built atop these binaries) and the choice of which to use themselves involve value judgements that can disadvantage certain people. People have intersectional identities and invariably belong to multiple groups simultaneously. When it comes to choosing fairness metrics, inherent to the majority of approaches in fair ML is that the researcher or practitioner decide what definition of fairness to use for their model. It has been shown that various definitions of fairness are not only mutually inconsistent but also prioritise different groups in different scenarios (Garg et al., 2020). In a sense then, solving for fairer ML models only pushes the problem from the model and onto the practitioner, as a 'fairer' model itself advantages and disadvantages different groups under different settings.

These two ethical considerations motivate the approach of our paper to conceptualise fairness in a more general setting where sensitive attributes can be continuous and multidimensional and fairer models are measured in terms of sensitive information removal. This conception avoids the ethical issues of binaries and fairness metrics.

We do note however that there still exist ethical concerns with our approach in terms of explainability. Measuring fairness by sensitive information removal (by measuring loss from a trained classifier) does not have an intuitive scale or unit of measurement for discussing the fairness or unfairness of a model. Although we can compare two models in terms of which is fairer, saying a model is fair because it scores some number in MSE has little intuitive meaning. Being unable to communicate the specifics of how a learned representation has removed sensitive information and how will affect downstream classifiers risks undermining confidence in fair ML as well.

Despite the explainability issue, we nonetheless believe that this approach represents a promising and exciting direction in fair ML that deal with substantive existing ethical issues. We hope that one area of future research may be deriving theoretical frameworks that can derive guarantees between sensitive information removal from debiased representations and upper bounds on downstream fairness metrics. This would develop a practical link to well-

known ideas of fairness and how unfair outcomes could appear in worst-case scenarios.