Improve Identity-Robustness for Face Models

Qi Qi [1] **Shervin Ardershir** [2]

Abstract

Despite the success of deep-learning models in many tasks, there have been concerns about such models learning shortcuts, and their lack of robustness to irrelevant confounders. When it comes to models directly trained on human faces, a sensitive confounder is that of human identities. Due to the privacy concern and cost of such annotations, improving identity-related robustness without the need for such annotations is of great importance. Here, we explore using off-the-shelf face-recognition embedding vectors, as proxies for identities, to enforce such robustness. Given an identity-independent classification task and a face dataset, we propose to use the structure in the face-recognition embedding space, to implicitly emphasize rare samples within each class. We do so by weighting samples according to their conditional inverse density (CID) in the proxy embedding space. Our experiments suggest that such a simple sample weighting scheme, not only improves the training robustness, it often improves the overall performance as a result of such robustness. We also show that employing such constraints during training results in models that are significantly less sensitive to different levels of bias in the dataset.

1. Introduction

Given the success of machine learning models, and their deployment at scale, having a more extensive evaluation of the robustness of such models is of utmost importance. Given the nature of training such models, there is always the potential for these models to rely on irrelevant and spurious shortcuts. Relying on such shortcuts could have immense negative consequences when the dataset and tasks are defined around humans. A prevalent type of such datasets and tasks are those defined on human faces, ranging from regression tasks such as estimating pose(Albiero et al., 2021), facial-landmarks(Wu & Ji, 2019), etc, to classification tasks such as facial-expressions classification(Huang et al., 2019), and generative tasks such as avatar creation(Alldieck et al., 2018), etc. A common attribute of many of such face-centric tasks is the fact that the model performance should be identity independent by definition. However, this aspect of a model is often not taken into account during training and evaluation. Two models trained on a face-related task can have similar overall performance, but very different levels of robustness across different individuals. The toy example in Figure 1 illustrates this concept. This disparity in performance often gets baked into the model due to bias in the training data, as data points belonging to different subpopulations may have a different level of class imbalance.

Awareness of identity/group labels would allow for mitigation approaches to prevent such bias, such as recent efforts in adversarial training (Zhang et al., 2018; Elazar & Goldberg, 2018), model interpretation method (Rieger et al., 2020) and objective regularization (Bechavod & Ligett, 2017), which aim to reduce the disparity between different groups using the ground-truth group labels $g \in G$. In many practical scenarios, however, such information is not available at scale during training and evaluation. Also, collecting such detailed annotation could be costly and undesirable due to three main reasons: First, annotating every sample data point with all their potential types of group-membership information could be extremely costly. Second, collecting and maintaining such detailed categorical labels on human faces raise data-privacy concerns. And third, the nature of many types of such group memberships may be extremely subjective. In addition to the previous hurdles in obtaining such data, most current large-scale datasets, lack such annotations at scale, which is another testament to the need for approaches that are not reliant on the availability of such additional information. As a result, improving fairness when the ground-truth group labels $g \in G$ are unknown is of utmost importance, and has given rise to an area of research often referred to as "fairness under unawareness". When it comes to "fairness under unawareness" for face models, the

[1]Department of Computer Science, The University of Iowa, Iowa City, IA, USA [2]Netflix, Los Gatos, CA, USA. Correspondence to: Qi Qi <qi-qi@uiowa.edu>, Shervin Ardershir <shervina@netflix.com>. The work was done while the first author was an intern at Netflix.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

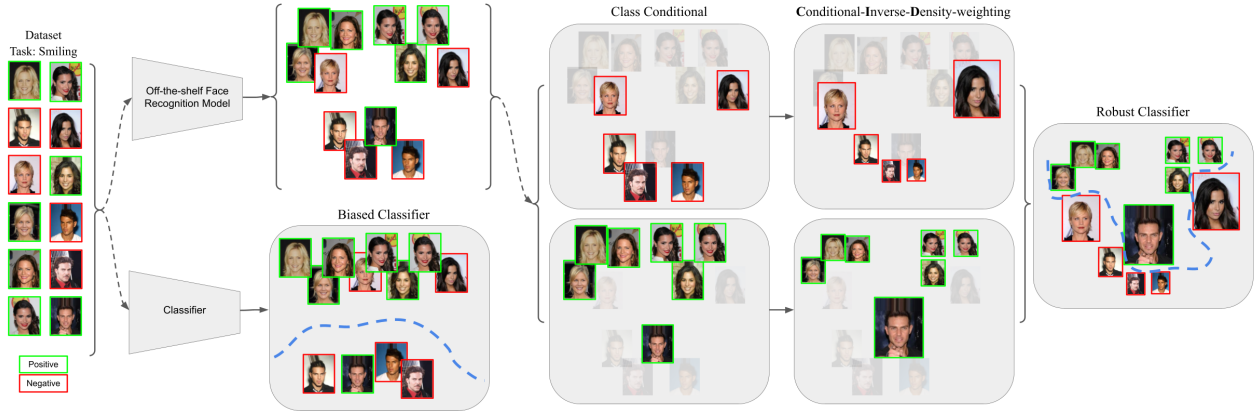

Figure 1. Toy example visualizing our proposed approach. The task is predicting if a face image is smiling (green) or not (red). The Biased Classifier shows how a biased dataset could lead to a model latching on to spurious features (identity) for an identity-independent task (smiling). We propose extracting face-recognition embeddings and using the structure in that space to weight rare samples within each class. More specifically, for each class (green or red), each sample is weighted based on its class-conditioned inverse density in the proxy (face recognition) embedding space. As a result, in each class, the rare samples are emphasized in the Robust Classifier.

only earlier work is (Ardeshir et al., 2022) which aims to measure the performance disparity of a model in the absence of group information. A disparity method (Disparity across Embedding Neighborhoods) is proposed, which approximates Rawlsian Max-Min (RMM) across groups $g \in G$, solely based on face-recognition embedding vectors. The neighbors of a sample are defined as the samples whose euclidean distance in the face-recognition embedding space is less than a predefined threshold. The aforementioned work solely focused on approximating disparity for a given model. In this work, however, we focus on using such intuition to reduce such disparity during training and directly optimize for such an objective. In other words, given a face dataset and solely its task labels, and without any group information, we explore if we can use embeddings from an off-the-shelf face recognition model to reduce the performance disparity of such a model across different individuals. Face Recognition models are often trained with a contrastive objective, and by pushing the samples belonging to the same person close to each other while pushing samples belonging to different individuals away from each other. As a result, the structure in the face recognition embedding space would ideally be variant with respect to facial features. Therefore, such embedding space would be a great proxy for facial features and thus identities, but in continuous space.

2. Related work

Our work can be categorized under the *Fairness under Unawareness* umbrella, where the aim is to enforce fairness across groups when the categorical group information G is unavailable during training. Given that most realistic scenarios of model training will have a similar setup, improving model robustness under such circumstances to minimize spurious correlations has become important to enhance model fairness. This can be achieved through various methods such as invariant risk minimization (Arjovsky et al., 2019; Adragna et al., 2020), distributionally robust optimization (Ben-Tal et al., 2013; Caton & Haas, 2020; Sagawa et al., 2019; Hashimoto et al., 2018; Li et al., 2021; Qi et al., 2020; Lahoti et al., 2020), and class balancing methods (Yan et al., 2020; Cui et al., 2019; Huang et al., 2016; Wang et al., 2017).

3. Approach

Given a dataset of images of faces, and an identity-independent face-related task such as predicting a facial expression (e.g. smiling), we aim to train a classifier that performs robustly across face images of different people. We refer to training labels related to the task of interest (smiling), as *task labels*. We assume that such labeling (whether a face is smiling or not) is given to us for training and test set. On the contrary, we assume that no *identity* label is given to us during training. Identity labels specify which images belong to which person (person-1, person-2, ...), across which we would like to enforce fairness/robustness. We also assume that we have access to an off-the-shelf face recognition model, using which we can extract an embedding for each face image. Our goal is to train a model for that task, that performs robustly (fairly) across different individuals on the test set. Please note that in our experiments, we solely use the identity-labeled test sets to validate the robustness of our approach, and we do not use such labels during training. Formally, given a dataset $\mathcal{D} = \{X \times Y\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ with size $n = |D|$, the total number of classes C, i.e, $|Y| = C$. $\mathcal{D}_y = \{(\mathbf{x}_i, y_i) | y_i = y, i \in [1, \cdots, |\mathcal{D}|]\}$ represents the

samples whose task label is $y \in Y$. $g_i \in G$ denotes the identity/group that sample i belongs to, across which performance disparity should be mitigated. Under our setup, group/identity labels G are unavailable during training. Instead, the embedding vectors $\{\mathbf{z}_i\}_{i=1}^{|\mathcal{D}|}$ are extracted from a face recognition model and are provided as proxies for the group/identity membership.

4. Training

Inspired by recent efforts in (Diana et al., 2021; Hashimoto et al., 2018; Lahoti et al., 2020), we define our objective as a min-max form, which encourages emphasis on the performance of the model on the least accurate areas of the embedding space:

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \frac{p_i^\tau}{Z_{y_i}} \ell(\mathbf{w}; \mathbf{x}_i, y_i) \tag{1}$$

$$\text{s.t} \arg \max_{\mathbf{p}_i \in \Delta_{\mathcal{D}_{y_i}}} \sum_{j \in \mathcal{D}_{y_i}} p_{ij} \mathbf{z}_i^\top \mathbf{z}_j - \tau \text{KL}(\mathbf{p}_i, \frac{\mathbf{1}}{|\mathcal{D}_{y_i}|}) \tag{2}$$

where $p_i^\tau := p_{ii}$ denotes the robust sample weight, $\ell(\mathbf{w}; \mathbf{x}_i, y_i)$ denotes the prediction loss, and $Z_{y_i} = \sum_{i \in \mathcal{D}_{y_i}} p_i^\tau$ is a class-level normalization parameter to guarantee each class contributes equally. The maximum constraint equation that obtains p_i^τ (2) is defined on the pairwise similarity of proxy embedding vectors, everaging the proxy neighborhood structure for each sample. To be more specific, for $\forall(\mathbf{x}_i, y_i) \sim \mathcal{D}$, $\mathbf{p}_i = (p_{i1}, \cdots, p_{ii}, \cdots, p_{i|\mathcal{D}_{y_i}|})$ refers to the importance weight assigned to each sample within the same class and satisfies $\Delta_{\mathcal{D}_{y_i}} := \{\sum_j p_{ij} = 1, p_{ij} \geq 0\}$. Maximize $\max_{\mathbf{p}} \sum p_{ij} \mathbf{z}_i^\top \mathbf{z}_j$ encourages the model to focus on itself, i.e, the close form solution of robust weight is $p_{ii} = 1, p_{ij} \in \mathbf{p}_i, j \neq i$ when $\tau = 0$. Hence, to explore the neighborhood structure in the proxy embedding space, we add the KL divergence regularizer $\sum_j p_{ij} \log(|\mathcal{D}_{y_i}| p_{ij})$ with $\tau > 0$ to penalize the discrepancy between $\mathbf{p}_i$ and the uniform weight $1/|\mathcal{D}_{y_i}|$, which encourages $\mathbf{p}_i$ focusing on the local neighborhood for any arbitrary $(\mathbf{x}_i, y_i) \sim \mathcal{D}$. The regularizer hyperparameter τ measures the proximity and magnitude of a neighborhood $(\mathbf{x}_i, y_i)$.

Due to the strong concavity of $\mathbf{p}_i$ in (2) and the specific structure of KL divergence, the close form solution of $p_i^\tau := p_{ii}$ is obtained by taking the first derivative of $\mathbf{p}$ in (2) equals to 0, i.e,

$$p_i^\tau = \frac{\exp(\frac{\mathbf{z}_i^\top \mathbf{z}_i}{\tau})}{\sum_{k=1}^{|\mathcal{B}_{y_i}|} \exp(\frac{\mathbf{z}_i^\top \mathbf{z}_k}{\tau})} \tag{3}$$

where the numerator is the exponential of the inner product of the proxy embedding vector $\mathbf{z}_i$ of sample $(\mathbf{x}_i, y_i)$. The denominator explores the neighborhood proxy structure by aggregating the exponential pairwise similarities of proxy vectors between sample $(\mathbf{x}_i, y_i)$ and $\mathcal{B}_{y_i}$. Hence, even though the constraint set is defined in $\mathcal{B}_{y_i}$, the skewness property of exponential function $\exp(\cdot/\tau)$ for large similarities pairs encourages the denominator to focus on the local neighbors of $(\mathbf{x}_i, y_i)$ that share the same facial features. $p_i^\tau \in (0, 1]$ represents the importance of the sample $(\mathbf{x}_i, y_i)$ in the local neighborhood. The fewer the samples in the local neighborhood, the higher the p_i^τ. Hence, p_i^τ is inversely proportional to the class-conditional sample density in the local neighborhood and emphasizes the rare samples within each class. This allows capturing a more nuanced notion of sample rarity within each class, which goes beyond the typical frequency-based methods

In (Ardeshir et al., 2022), the performance of a model across different local neighborhoods in the proxy embedding space is used to estimate disparity across identities/groups. Hence a local neighborhood could be seen as an approximation for a subpopulation/group/identity g_i. (Ardeshir et al., 2022) also illustrates that there are different neighborhood sizes that better approximate different group memberships. To capture the same concept, in our formulation τ controls the skewness of the exponential function, which influences the size of the local neighborhood. Thus, we fine-tune the hyperparameter τ to allow for exploring different neighborhood sizes and therefore different density estimations. As it can be seen, as $\tau \to \infty$, the weights converge to $p_i^\tau \to \frac{1}{|\mathcal{B}_{y_i}|}$ which is simply the inverse of per-class frequency.

4.1. Area Under Min-Max Curves (AUMM)

As mentioned earlier, we do not have access to group labels during training, however, to measure if our model is in fact more robust across groups, we use group labels in the test set to validate our hypothesis. In our setup, we mostly focus on robustness/fairness across individuals, and given that the number of individuals in a face dataset could be very large, we define a modification to the widely used Rawlsian min-max metric. In the Rawlsian min max metric (Rawls, 2001), the ratio of the performance of the model is measured between the most and least accurate groups, i.e. $1 - \frac{\min_g(e_g)}{\max_g(e_g)}|_{g \in G}$. This measure is often very useful when the number of groups is very limited. However, given that in our instance, we are interested in measuring disparity across different people, the number of different individuals in the dataset could be very large. Therefore, using the ratio of performance only on the highest and lowest individual will ignore large portions of the dataset. Thus we modify the Rawlsian min-max formulation to measure the ratio of the bottom-k% and top-k% of groups instead.

$$\text{MM} = \{1 - \frac{\bar{e}_g^k}{\underline{e}_g^k}\}_{k=1}^{|G|} \tag{4}$$

where $k \in [1, \cdots, |G|]$ denotes the index of groups. $\bar{e}_g^k$ and $\underline{e}_g^k$ are the average of top and bottom k group performance, respectively. Sweeping k results in a curve, which we refer

Algorithm 1 CID Optimization (τ)

Model initialization $\mathbf{w}_1$, proxy embeddings $\{\mathbf{z}_i\}_{i=1}^n$
for $t = 1, \ldots, T$ **do**
  Sample a batch of B samples $\mathcal{B} = \{(\mathbf{x}_i, y_i)\}_i^B \sim \mathcal{D}$
  Retrieve the proxy embedding vectors of batch samples, $\{\mathbf{z}_i\}_{i=1}^B$.
  Calculate p_i^τ according to Eqn (3) for $\forall(\mathbf{x}_i, y_i) \in \mathcal{B}$
  Calculate $Z_{y_i} = \sum_{j \in \mathcal{B}_{y_i}} p_j^\tau$
  Calculate CID loss: $\sum_{i \in \mathcal{B}} p_i^\tau \ell_i(\mathbf{w}_t)/(Z_{y_i})$
  Update $\mathbf{w}_t$ using stochastic algorithms.
end for
Return $\mathbf{w}_{T+1}$,

to as the Min Max Curve. We use the area under this curve, AUMM for short, as a metric for robustness across groups. The lower the AUMM, the more robust/fair the model is.

5. Experiments

Baselines: We compare the proposed approach (CID) with four baselines on fairness under awareness setup: **IFW** (inverse frequency weighting) (Huang et al., 2016; Wang et al., 2017), **DRO** (Distributionally Robust Optimization)(Li et al., 2021; Qi et al., 2020), **IRM** (Invariant Risk Minimization) (Adragna et al., 2020), and **ARL** (Adversarial Reweighted Learning)(Lahoti et al., 2020). The details of the methods are provided in the appendix.

Dataset: CelebA (Liu et al., 2015) has 200K face images and includes 10117 identities in total. Each image is labeled with 40 attributes/tasks. We pick identity-independent task (Liu et al., 2015). *Smiling* and train standard binary classification models to predict those tests. We extract its face-recognition embedding vector $\mathbf{z}$ using the face recognition model (King.) and use it as its identity proxy.

Evaluation Metric We evaluate the performance of the baselines, in terms of overall classification accuracy (Acc), average, and standard deviation of per-identity accuracy (Id Acc and δ_{Id}). A low δ_{Id} is one of the metrics implying that the performance of the model is robust across identities and thus fairer. Also, we use the area under the min-max curve (AUMM for short) described in section 4.1 as another robustness metric. Given that this metric measures the disparity between the accuracy of the top-k and bottom-k identities, the lower the AUMM, the more identity-robust a model is.

Stress-testing with Controlled Bias We conduct a set of experiments, in which we explicitly control the train dataset bias and measure how the model's performance sensitivity to dataset bias. To do so, we construct different versions of the train sets of CelebA, by manipulating the dataset and adding controlled artificial identity-to-task bias. More specifically, given a task (such as smiling), we construct a biased train set by excluding $p\%$ of data points belonging to a (task label, subpopulation). If we train a classifier on such a dataset, a model's performance on the standard (non-manipulated) test set can be very non-robust, as the train and test set do not follow the same distribution.

In all the setups we solely manipulate the bias in the train set and keep the test set unchanged. We try this experiment for different values of $p \in \{25\%, 50\%, 75\%, 90\%\}$, and for different (group, task-label) combinations. We refer to each setting by specifying which group (M:Male vs. F:Female), and which task label (P: positive, N: negative) has been manipulated (excluded by p%) from the training and validation set (while the test set is unchanged). As an example, on the task "Smiling", FP 50% means that half of the Female Positives (smiling female faces) were excluded during training, therefore creating a bias in the dataset.

Table 1 and Figure 2 show the results of our stress test on the task "Smiling". The proposed CID method has smaller MMC, AUMM, and Id values, in addition to higher bottom 10% Id accuracy compared to the baselines. When it comes to the level of bias, the gap between the models steadily increases as the amount of induced bias in the dataset increases, which verifies the advantages of CID over CE on handling distribution shift.[1]

6. Conclusion

Our experiments show that our simple sample-weighting approach helps face models to maintain high accuracy while gaining significant robustness to distribution shifts and different levels of bias, and often maintaining a more uniform performance across different identities (and groups) of faces without the need for group labels during training.

Table 1. Stress testing the models on the CelebA dataset, by eliminating 90% of a subpopulation in the training/validation set.

FP	Acc	Id Acc	10% Id Acc	δ_{Id}	AUMM
CE	90.33	89.77	65.91	0.1095	0.1821
IFW	91.13	90.49	67.28	0.1053	0.1737
DRO	90.23	89.84	66.20	0.1098	0.1834
IRM	91.18	90.11	66.54	0.1088	0.1789
ARL	90.40	89.46	65.55	0.1127	0.1863
CID	**91.31**	**90.67**	**67.73**	**0.1042**	**0.1717**
MN	Acc	Id Acc	10% Id Acc	δ_{Id}	AUMM
CE	90.29	89.53	63.72	0.1151	0.1922
IFW	90.07	90.24	65.68	0.1093	0.1816
DRO	90.03	89.60	63.91	0.1146	0.1915
IRM	90.06	89.70	64.11	0.1123	0.1845
ARL	90.33	89.51	63.54	0.1162	0.1956
CID	**91.20**	**90.43**	**66.16**	**0.1080**	**0.1783**

[1]Due to space limitations, we only provide a subset of setup experiments in the paper. For more results, including more datasets, please visit our full paper.

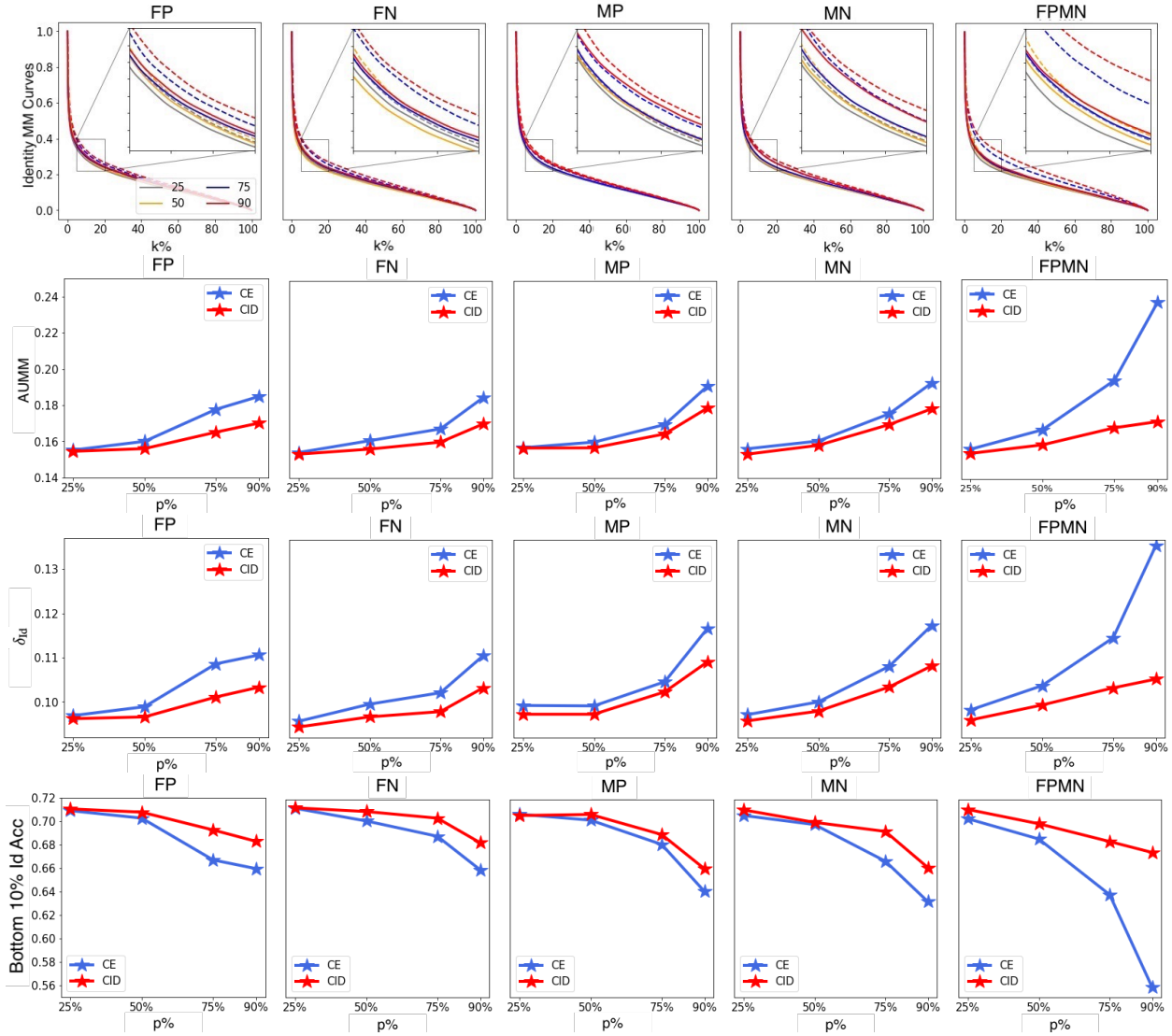

Figure 2. Results of different stress tests on the CelebA dataset for the smiling task label. First row: For the MM figure, the x-axis shows k for which the disparity between top and bottom k identities is evaluated. the In each figure, the x-axis specifies the amount (percentage) of the training data of the (group, task label) that is excluded during training. As can be observed, CID maintains its original metrics significantly better than CE in the presence of a distribution shift.

References

Adragna, R., Creager, E., Madras, D., and Zemel, R. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*, 2020.

Albiero, V., Chen, X., Yin, X., Pang, G., and Hassner, T. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7617–7627, 2021.

Alldieck, T., Magnor, M., Xu, W., Theobalt, C., and Pons-Moll, G. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pp. 98–109. IEEE, 2018.

Ardeshir, S., Segalin, C., and Kallus, N. Estimating structural disparities for face models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10358–10367, 2022.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bechavod, Y. and Ligett, K. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Caton, S. and Haas, C. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 66–76, 2021.

Elazar, Y. and Goldberg, Y. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640*, 2018.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5375–5384, 2016.

Huang, Y., Chen, F., Lv, S., and Wang, X. Facial expression recognition: A survey. *Symmetry*, 11(10):1189, 2019.

King., D. https://github.com/davisking/dlib-models.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. On tilted losses in machine learning: Theory and applications. *arXiv preprint arXiv:2109.06141*, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Qi, Q., Xu, Y., Jin, R., Yin, W., and Yang, T. Attentional biased stochastic gradient for imbalanced classification. *arXiv preprint arXiv:2012.06951*, 2020.

Rawls, J. *Justice as fairness: A restatement*. Harvard University Press, 2001.

Rieger, L., Singh, C., Murdoch, W., and Yu, B. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International conference on machine learning*, pp. 8116–8126. PMLR, 2020.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Wang, Y.-X., Ramanan, D., and Hebert, M. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

Wu, Y. and Ji, Q. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127 (2):115–142, 2019.

Yan, S., Kao, H.-t., and Ferrara, E. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1715–1724, 2020.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

A. You *can* have an appendix here.

You can have as much text here as you want. The main body must be at most 4 pages long. For the final version, one more page can be added. If you want, you can use an appendix like this one, even using the one-column format.