

Invariant Feature Regularization for Fair Face Recognition

Jiali Ma¹ Zhongqi Yue² Kagaya Tomoyuki³ Suzuki Tomoki³
Karlekar Jayashree¹ Sugiri Pranata¹ Hanwang Zhang²

¹Panasonic R&D Center Singapore ²Nanyang Technological University ³Panasonic Connect Co., Ltd. R&D Division
jiali.ma@sg.panasonic.com yuez0003@ntu.edu.sg kagaya.tomoyuki@jp.panasonic.com
suzuki.tomoki@jp.panasonic.com karlekar.jayashree@sg.panasonic.com
sugiri.pranata@sg.panasonic.com hanwangzhang@ntu.edu.sg

Abstract

Fair face recognition is all about learning invariant feature that generalizes to unseen faces in any demographic group. Unfortunately, face datasets inevitably capture the imbalanced demographic attributes that are ubiquitous in real-world observations, and the model learns biased feature that generalizes poorly in the minority group. We point out that the bias arises due to the confounding demographic attributes, which mislead the model to capture the spurious demographic-specific feature. The confounding effect can only be removed by causal intervention, which requires the confounder annotations. However, such annotations can be prohibitively expensive due to the diversity of the demographic attributes. To tackle this, we propose to generate diverse data partitions iteratively in an unsupervised fashion. Each data partition acts as a self-annotated confounder, enabling our Invariant Feature Regularization (INV-REG) to deconfound. INV-REG is orthogonal to existing methods, and combining INV-REG with two strong baselines (Arcface and CIFP) leads to new state-of-the-art that improves face recognition on a variety of demographic groups. Code is available at <https://github.com/milliema/InvReg>.

1. Introduction

Face recognition is essentially an out-of-distribution generalization problem, where the goal is to learn feature that generalizes to unseen faces in deployment [32]. In particular, due to its wide application in sensitive areas such as crime prevention [41], fair face recognition becomes a pressing need [10, 14, 35]. This means that the model must perform equally well on all demographic groups, *i.e.*, it learns the causal feature (*e.g.*, face identity) invariant to the demographic attributes (*e.g.*, race or gender).

However, demographic attributes are naturally imbalanced in data at scale, *e.g.*, as shown in Figure 1 first row, in the prevailing MS-Celeb-1M dataset [12], “non-Caucasian” and “female” are the minority racial and gender groups, re-

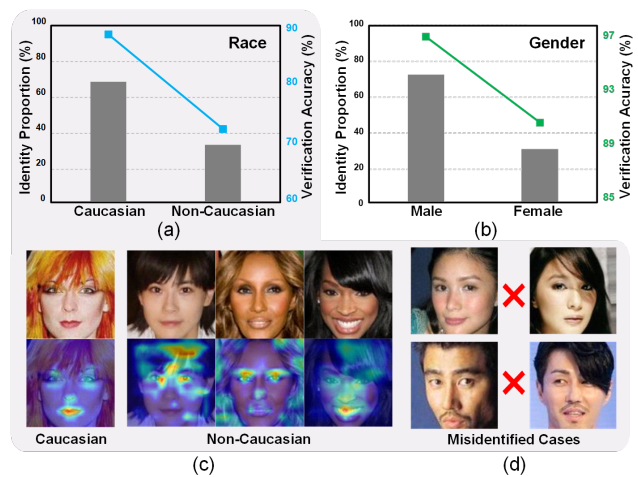


Figure 1: Biased model trained on the dataset with imbalanced demographic attributes. (a) and (b) show the attribute proportions and verification accuracy on race and gender, respectively. (c) Grad-CAM [31] attention maps of face images from two racial groups. (d) Misidentified cases.

spectively. Furthermore, a model naively trained on this imbalanced dataset underperforms on the minority groups in deployment, *e.g.*, more than 10% accuracy decrease on “non-Caucasian”, and 5% decrease on “female”. In particular, by analyzing the attention maps across the two racial groups in Figure 1c, we observe that besides the causal feature (*e.g.*, facial attributes like eyes and nose), the model additionally focuses on the spurious demographic-specific feature (*e.g.*, hairstyle) on the minority group. This is because the less diverse face images in each non-Caucasian identity tend to share the hairstyle, making it a valid context to distinguish identities. Yet this demographic-specific context generalizes poorly to unseen faces, *e.g.*, misidentifying the same identity with different hairstyles as in Figure 1d.

From a causal point of view, the bias stems from the confounding effect [26, 27, 11]. In face recognition, when trained to predict the identity label Y given the face image

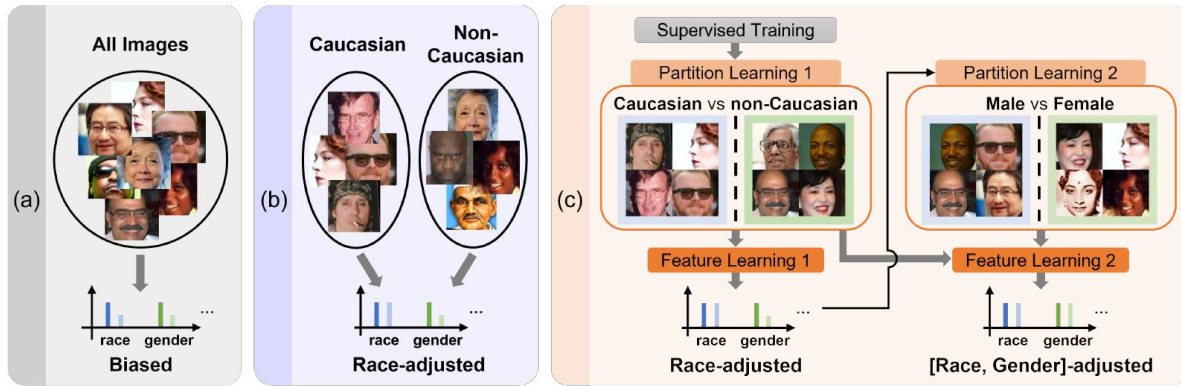


Figure 2: Comparison of different face recognition approaches. (a) Standard supervised training (biased to race and gender). (b) Learning with ground-truth race partition (biased to gender). (c) Our INV-REG without using any annotation of ground-truth demographic attribute (invariant to race and gender).

X , the model is confounded by the demographic attributes D (e.g., $D = \text{“non-Caucasian”}$), which is the common cause of X and Y . Specifically, the training image X is sampled from the demographic group specified by the attributes D (i.e., $D \rightarrow X$), and any demographic-specific feature that is discriminative towards Y (e.g., hairstyle) serves as contextual cue when predicting the identity (i.e., $D \rightarrow Y$). Hence, in pursuit of a lower classification loss, the model recklessly exploits the context feature ($X \leftarrow D \rightarrow Y$) which may not generalize to unseen faces. The confounding effect is more pronounced on the minority group, e.g., in the extreme case where $D = \text{“dark skin”}$ only has one identity, learning the “dark skin” feature alone ($D \rightarrow Y$) can tell the identity apart from “light skin” ones, yet this spurious feature fails to distinguish different “dark skin” identities in deployment.

The aforementioned confounding effect can only be removed by causal intervention [27, 26]. One way is to collect a balanced dataset with diverse faces in all demographic groups (i.e., adjusting the demographic attribute D). After all, if there is no demographic-specific context in any group, the model can only learn the causal feature. However, this is impractical due to the prohibitive data collection cost. Note that “balanced” without “diverse” is ineffective since the spurious context still persists, as shown by the limited success on small balanced dataset [34]. The other way is backdoor adjustment [26], which can be implemented by data partition (Section 3.2). For example, in contrast to the naive supervised training on all images in Figure 2a, some works [11, 34, 19, 35] partition the training images into “race” splits (i.e., adjusting “race”), and train a model invariant across the splits as illustrated in Figure 2b. However, since the demographic attributes are diverse in practice, using only the “race” splits is far from sufficient.

To this end, we propose Invariant Feature Regularization, dubbed as **INV-REG**, which iteratively self-annotates the confounders by learning data partitions, as illustrated in Figure 2c. Our INV-REG hinges on the invariance of causal

relation [26, 25]: causal feature is invariantly discriminative across the splits in any confounder partition. Its contra-position enables **1) partition learning** about confounder: if the current feature is not invariantly discriminative across the learned splits, the partition corresponds to a confounding demographic attribute. Then we perform **2) feature learning** to achieve invariance across the learned splits (i.e., causal intervention), which removes its confounding effect. We iterate between the two steps to learn causal feature invariant to diverse demographic attributes.

Our contributions are summarized below:

- We propose a partition learning strategy to self-annotate the demographic attributes (i.e., confounder) in the form of data partitions (Section 4.1). In particular, our approach discovers diverse demographic partitions without relying on any ground-truth annotation.
- We use the discovered partitions to impose an invariant regularization in training to learn causal feature robust in all demographic groups (Section 4.2).
- Overall, INV-REG is a regularization module orthogonal to existing face recognition methods. Combining INV-REG with two strong baselines leads to new state-of-the-art results, i.e., 79.44% on Arcface and 81.17% on CIFP for average multi-racial accuracy (Section 5.2).

2. Related Work

Fair Face Recognition. Conventional face recognition explores different loss functions for improved feature learning. For example, CenterLoss [39] and RangeLoss [46] penalize the distance between feature and the corresponding class center. SpheroFace [21], Cosface [33], and Arcface [8] leverage margin loss to promote intra-class density and inter-class separability. However, they are easily biased towards the majority group in the training data. In pursuit of fairness, some works construct balanced datasets w.r.t. race and gender attributes at limited scale [34, 14, 28, 35]. Other

works leverage group-level (*e.g.*, race) [34, 19] or sample-level margins [43, 4, 18, 9, 38, 42] for re-weighted training, such that minority groups or hard samples (*e.g.*, misclassified training samples) are assigned a larger training weight. In contrast, our INV-REG grounds the elusive confounding demographic attributes into concrete data partitions as self-annotated confounders and performs causal intervention to explicitly remove the bias, leading to state-of-the-art accuracy across a variety of demographic groups.

Partition-Based Learning is an effective tool towards out-of-distribution generalization. Some works [29, 45] divide the training data into attribute-based groups and optimize the worst-group loss. Other works [30, 3, 44] down-weight or sub-sample the majority group to artificially balance the confounding attributes. In face recognition, a few works [35, 34] partition data by race and transfer knowledge from the majority “Caucasian” to other races. However, they rely on ground-truth race annotations and lack diversity in practice due to the expensive labeling cost. Some works also leverage self-annotated partitions, *e.g.*, to improve the attention module [37], learn a disentangled representation [36], or remove spurious background feature [6]. In contrast, our INV-REG is tailored for fair face recognition and critically differs in 1) assigning partitions at identity-level instead of image-level to remove spurious demographic attributes defined across identities; 2) performing iterative partition learning and feature learning steps to progressively address various spurious features (*e.g.*, race, gender, *etc.*).

3. Problem Formulation

3.1. Face Recognition

Data and Model. We denote training data as $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where \mathbf{x}_i is an image and $y_i \in \{1, \dots, C\}$ denotes its identity label among C different identities. We drop the subscript i for simplicity when the context is clear. The demographic attribute labels are generally not available in training. The model consists of a feature backbone $\Phi(\mathbf{x})$ that outputs a feature vector, and a classifier f that outputs the prediction logit for each of the C identities given a feature.

Train and Test. In training, Φ and f are optimized by classification loss (*e.g.*, cross-entropy) for identity prediction. In testing, Φ is evaluated on unseen samples $\{\mathbf{x}_i, y_i, \mathbf{d}_i\}_{i=1}^M$ where \mathbf{d}_i is a tuple of demographic attributes, *e.g.*, $\mathbf{d}_i = (\text{race: “Caucasian”, gender: “Male”})$. A fair face recognition model should learn feature $\Phi(\mathbf{x})$ that is discriminative towards y in each demographic group (*e.g.*, “Caucasian”, “Male”). The exact evaluation protocol is in Section 5.1.

Causal View. We depict the causal relations in face recognition with a causal graph in Figure 3. $X \rightarrow Y$ represents the desired causal effect from an image X to its identity Y , D denotes the set of *all* demographic attributes and is diverse in practice (*e.g.*, CelebA in Section 5.2).

$D \rightarrow X$ denotes that the attributes affect the appearance of image X (*e.g.*, few non-Caucasians have light skin). In particular, $D \rightarrow Y$ is because D contains demographic-specific contextual cue for predicting Y due to dataset bias (*e.g.*, hairstyle is discriminative due to limited non-Caucasian training images in Figure 1c). $X \leftarrow D \rightarrow Y$ is known as the backdoor path, which misleads the model to capture the spurious confounding effect [26, 27] (*e.g.*, predicting with hairstyle in Figure 1d). This confounding effect can only be removed by causal intervention [27], and we introduce an effective implementation below called invariant learning.

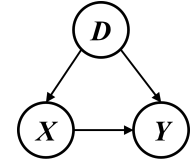


Figure 3: Causal graph of face recognition.

3.2. Invariant Learning

Given a partition of the training dataset into subsets and a loss function, the goal is to learn a model that simultaneously minimizes the loss in all the subsets of the partition.

Partition. In particular, we specify a partition of the C training identities into K subsets by the partition matrix $\mathbf{P} \in \{0, 1\}^{C \times K}$, where $P_{y,k} = 1$ if the y -th identity belongs to the k -th subset, and 0 otherwise.

Loss. We are interested in supervised classification loss in face recognition. Given a feature extractor Φ and classifier f , we denote the loss in the k -th subset of \mathbf{P} as $\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k)$.

Objective. Invariant learning uses the following objective, called the invariant risk minimization (IRM) [1]:

$$\begin{aligned} \min_{\Phi, f} \sum_{k=1}^K \mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k), \\ \text{s.t. } f \in \bigcap_{k=1}^K \underset{f^*}{\operatorname{argmin}} \mathcal{L}_{cls}(\Phi, f^*, \mathbf{P}, k). \end{aligned} \quad (1)$$

The constrained optimization minimizes the combined classification loss (first line) while maintaining an invariant classifier f by keeping the losses in all subsets simultaneously optimal (second line). This objective learns the causal feature invariant across the subsets. For example, in Figure 1c, while exploiting “hairstyle” can reduce \mathcal{L}_{cls} in the minority “non-Caucasian”, it hurts the loss in “Caucasian” as face images of the same identity may present different hairstyles. Therefore, the spurious “hairstyle” feature will be removed to satisfy the invariance constraint.

To avoid the challenging bi-level optimization in Eq. (1), we use two practical implementations—IRMv1 [1] and REX [17], which merge the classification loss and the con-

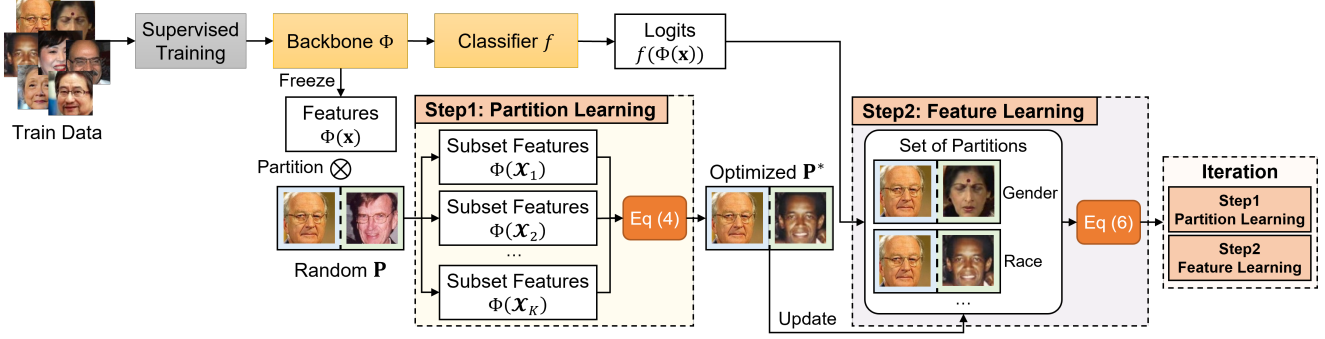


Figure 4: Pipeline of INV-REG. After standard supervised training for 1 epoch, we first conduct [Step 1 Partition Learning] to discover a partition corresponding to a confounding demographic attribute (Section 4.1). Next, we perform [Step 2 Feature Learning] to achieve invariance across the partition subsets (Section 4.2). We iterate the above two steps until convergence.

straint as a single invariant loss, denoted as $\mathcal{L}_{inv}(\Phi, f, \mathbf{P})$:

$$\mathcal{L}_{inv}(\Phi, f, \mathbf{P}) = \sum_{k=1}^K \mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k) \quad \text{IRMv1 [1]} \\ + \lambda \|\nabla_{\mathbf{w}=1.0} \mathcal{L}_{cls}(f \circ \Phi, \mathbf{w}, \mathbf{P}, k)\|_2^2 \quad (2)$$

$$\mathcal{L}_{inv}(\Phi, f, \mathbf{P}) = \sum_{k=1}^K \mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k) \quad \text{REx [17]} \\ + \lambda \text{Var}(\{\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k)\}_{k=1}^K), \quad (3)$$

where λ is the invariant penalty strength, and $\text{Var}(\cdot)$ computes the variance of a set of losses. Intuitively, the second term in IRMv1 [1] treats the logits produced by the composition $f \circ \Phi$ as a feature and computes the optimality of a baseline classifier $\mathbf{w} = \mathbf{1.0}$. Hence IRMv1 achieves invariance by encouraging a fixed baseline to be simultaneously optimal in all subsets. In contrast, REX minimizes the variance of all subset losses to achieve invariance. We discuss the choice of \mathcal{L}_{inv} and penalty strength λ in Section 5.4.

4. Proposed Method

The overall pipeline of our proposed INV-REG is shown in Figure 4 and summarized in Algorithm 1. We first conduct standard supervised training (*e.g.*, using Arcface) for one epoch. Then we iterate between the following steps until convergence: 1) In partition learning, we fix the model and learn a partition matrix \mathbf{P} that *maximizes* \mathcal{L}_{inv} to discover a confounding demographic attribute (Section 4.1). 2) We perform feature learning by *minimizing* \mathcal{L}_{inv} to achieve invariance across all discovered partitions (Section 4.2). Note that our implementation of the two steps is simple, *i.e.*, choosing an appropriate form of the classification loss \mathcal{L}_{cls} to compute \mathcal{L}_{inv} . We detail the two steps below.

4.1. Partition Learning

The goal is to find a partition of the training dataset such that the current feature is not invariantly discriminative

across the partition subsets. For example in Figure 2c, the race-adjusted model is not invariant across the gender partition, *i.e.*, it is still confounded by the gender attribute, which should be addressed in the subsequent learning. Specifically, this goal corresponds to finding a partition matrix \mathbf{P}^* that maximizes \mathcal{L}_{inv} (*i.e.*, not invariant), while keeping the feature frozen (*i.e.*, current feature)¹:

$$\mathbf{P}^* = \underset{\mathbf{P}}{\text{argmax}} \mathcal{L}_{inv}(\Phi, f, \mathbf{P}). \quad (4)$$

In particular, when computing $\mathcal{L}_{inv}(\Phi, f, \mathbf{P})$, we adopt the supervised contrastive loss [16] as \mathcal{L}_{cls} to evaluate the feature in each subset of \mathbf{P} , as formulated below:

$$\mathcal{L}_{cls}(\Phi, \cdot, \mathbf{P}, k) = \sum_{\mathbf{x} \in \mathcal{X}_k} \sum_{\mathbf{x}^* \in \mathcal{X}_k} -\log \frac{e^{\Phi(\mathbf{x})^\top \Phi(\mathbf{x}^*)}}{\sum_{\mathbf{x}^* \neq \mathbf{x}} e^{\Phi(\mathbf{x})^\top \Phi(\mathbf{x}^*)}}, \quad (5)$$

where \cdot denotes that the classifier f is not used to compute the loss, $\mathcal{X}_k = \{\mathbf{x} | \mathbf{P}_{y,k} = 1\}$ is the set of training images in the k -th subset of \mathbf{P} , and \mathbf{x}^+ is an image sharing identity with \mathbf{x} in \mathcal{X}_k , known as the positive sample in contrastive learning. Notice that we do not use f in partition learning. The intuition is that, face recognition hinges on learning invariantly discriminative *feature* towards identity. Hence, we seek for confounders (in the form of partitions) that undermine such invariance on *feature* level. Our adopted supervised contrastive loss measures if the features from the same identity are clustered, and those from different identities are pushed away, making it a favorable choice in this regard.

Overall, we maintain a set of partitions \mathcal{P} that contains all the discovered partitions. \mathcal{P} is initialized as an empty set before training. After finding each \mathbf{P}^* with Eq. (4), we update it with $\mathcal{P} \leftarrow \mathcal{P} \cup \{\mathbf{P}^*\}$ for subsequent feature learning.

4.2. Feature Learning

The goal is to learn invariant feature that is simultaneously discriminative across the subsets of all the discovered

¹We optimize a continuous partition matrix in $\mathbb{R}^{C \times K}$ to enable back-propagation and threshold it to $\{0, 1\}^{C \times K}$. Details in Appendix.

Algorithm 1 INV-REG Training

- 1: **Input:** training images $\{\mathbf{x}_i, y_i\}_{i=1}^N$, randomly initialized Φ and f , empty set $\mathcal{P} = \{\}$
 - 2: **Output:** trained Φ
 - 3: Train Φ, f on $\{\mathbf{x}_i, y_i\}_{i=1}^N$ with no partition for 1 epoch
 - 4: **repeat**
 - 5: # **Step 1: Partition Learning (Section 4.1)**
 - 6: Compute $\mathcal{L}_{cls}(\Phi, \cdot, \mathbf{P}, k) \forall k \in \{1 \dots K\}$ with Eq. (5)
 - 7: Freeze Φ, f , learn \mathbf{P}^* with Eq. (4)
 - 8: Update $\mathcal{P} \leftarrow \mathcal{P} \cup \{\mathbf{P}^*\}$
 - 9: # **Step 2: Feature Learning (Section 4.2)**
 - 10: Compute $\mathcal{L}_{cls}(\Phi, f, \mathbf{P}, k), \forall k \in \{1 \dots K\}$
 - 11: Freeze \mathcal{P} , learn Φ and f with Eq. (6)
 - 12: **until** convergence
-

partitions in \mathcal{P} . For example in Figure 2c, by learning with the racial and gender partitions, the model becomes invariant to both demographic attributes. Hence, we optimize the model Φ and f by minimizing \mathcal{L}_{inv} :

$$\min_{\Phi, f} \sum_{\mathbf{P} \in \mathcal{P}} \mathcal{L}_{inv}(\Phi, f, \mathbf{P}), \quad (6)$$

where each $\mathbf{P} \in \mathcal{P}$ is frozen. As the goal is feature learning, we can plug in the classification loss in existing face recognition methods as \mathcal{L}_{cls} to compute \mathcal{L}_{inv} . In this paper, we choose two strong baselines: Arcface [8] and CIFP [43]. In a nutshell, Arcface increases the decision boundary in angular space such that features of different identities are separated by a user-defined angular margin at least. CIFP uses an adaptive margin for sample re-weighting, where hard samples (*i.e.*, misclassified in training) are assigned a stringent margin to increase the training weight. We include the detailed form of the two losses in Appendix.

5. Experiments

5.1. Experimental Setting

Training Dataset. We adopted the refined MS-Celeb-1M dataset [12, 8]. It is a representative large-scale dataset consisting of 85K celebrity identities with 5.8M images. It exhibits demographic imbalance, where ‘‘Caucasian’’ and ‘‘Male’’ are the majority groups as shown in Figure 1. Pre-trained attribute classification models are utilized for the demographic statistical analysis, details are in Appendix.

Testing Dataset. We performed extensive experiments on various benchmarks with different emphases, summarized in Table 1. We evaluated fair face recognition on 3 datasets with attribute labels: (1) **MFR** [7] is a large-scale non-celebrity dataset with racial attribute annotations. It serves as a challenging benchmark to test model generalization in different racial groups. (2) **RFW** [35] is a race-balanced dataset constructed from the original MS-Celeb-1M [12].

Dataset	Description	#Identity	#Image
MS-Celeb-1M [12]	Train set, imbalanced	85K	5.8M
MFR [7]	Multi-race, large-scale	242K	1.6M
RFW [35]	Multi-race	11K	40K
CelebA [22]	40 fine-grain attributes	40K	1985
IJB-B [40]	Images, video frames	1845	76.8K
IJB-C [23]		3531	148.8K
DigiFace-1M [2]	Large variations, synthetic faces	1000	72K

Table 1: Statistics of train and test data.

We followed [19] to exclude the overlapping identities from RFW with the training dataset. (3) **CelebA** [22] provides ground-truth labels for fine-grained attributes *e.g.*, gender, hair color and style. We combined data from validation and test splits to evaluate our models. In addition, we used 2 datasets to evaluate conventional face recognition. (4) **IJB-B** and **IJB-C** [40, 23] are large-scale datasets with images and continuous video frames collected under unconstrained scenarios. (5) **DigiFace-1M** [2] is the latest synthetic dataset of realistic face images. It renders each identity under large variations, *e.g.*, hair density, makeup, face-wear, and expression. Hence we adopted it to challenge the generalization ability of models. We used the first 1K identities for testing, where each identity has 72 images.

Implementation Details. We used image crops with the size of 112×112 as per standard [20]. We adopted the modified ResNet-50 and ResNet-100 in [8] as the backbones. We trained the models with SGD, using batch size of 512 and 21 total epochs. All the experiments were conducted on 4 NVIDIA Tesla V100 GPU with Pytorch framework. For the evaluation protocol, we reported the True Positive Rate (TPR) at a fixed False Positive Rate (FPR). We used FPR of $1e-4$ for both IJB-B and IJB-C datasets and followed common settings for the rest if not specifically mentioned.

5.2. Fair Face Recognition

MFR. In Table 2, we observe clear improvements on *all* races by plugging our INV-REG into Arcface and CIFP. For example on ResNet-50, we outperform Arcface baseline by 2.46% on African, 1.41% on South Asian, and 1.66% on East Asian, and we even improve the majority race by 0.87% on Caucasian. Furthermore, our method reduces the standard deviation over both baselines, demonstrating more balanced performance across the demographic groups. Note that the low standard deviation of Anchorface on ResNet-100 is because it sacrifices the accuracy of Caucasian. Overall, we have the following observations: 1) INV-REG greatly enhances the performance of minority groups without sacrificing that of the majority one, leading to the best ‘‘Avg’’ and ‘‘All’’ performance. This validates the effectiveness of our invariant learning in removing the confounding effects and capturing causal feature. 2) Our improvement on CIFP

	Method	African (AF)	Caucasian (CA)	South Asian (SA)	East Asian (EA)	Avg	Std	All
ResNet-50	Arcface* [8]	74.54	84.43	81.47	53.27	73.43	12.18	77.91
	Ours-Arcface	77.00	85.30	82.88	54.93	75.03	11.99	78.98
	CIFP* [43]	77.26	85.52	83.76	55.74	75.57	11.86	80.34
	Ours-CIFP	79.41	86.53	84.99	57.82	77.19	11.49	81.37
ResNet-100	Anchorface [20]	79.31	87.00	85.59	59.70	77.90	10.90	82.06
	Arcface [†] [8]	79.12	87.18	85.50	55.81	76.90	12.54	80.73
	Ours-Arcface	81.76	89.16	87.64	59.20	79.44	12.01	82.85
	CIFP* [43]	82.55	89.40	88.77	60.58	80.32	11.71	84.36
	Ours-CIFP	83.70	90.04	89.01	61.91	81.17	11.38	84.73

Table 2: Verification performance (%) on MFR dataset. (“*”: self-implemented results based on the officially released code. “†”: tested results using the released model from the author. “Ours-”: our results achieved by plugging our INV-REG into other baselines. “Avg”/“Std”: average/standard deviation of the accuracy on four races. “All”: accuracy on all the samples.)

Method	AF	CA	SA	EA	Avg	Std
Arcface [8]	97.48	98.80	97.38	96.80	97.61	0.73
LDAM-Cosface [5]	97.80	98.93	97.50	97.23	97.86	0.65
MetaCW [15]	97.86	99.13	98.11	97.73	98.20	0.55
MvCoM-URFace [19]	97.18	98.85	96.98	97.15	97.54	0.76
MvCoM-Cosface [19]	98.06	99.16	98.28	97.78	98.32	0.51
Ours-Arcface	98.56	99.47	98.76	98.39	98.79	0.41

Table 3: Verification accuracy (%) on RFW (ResNet-100).

Method	Male		Female	
	FPR=1e-5	FPR=1e-4	FPR=1e-5	FPR=1e-4
Arcface [8]	96.35	97.14	90.86	93.92
Ours	96.61	97.30	91.66	94.17

Table 4: Accuracy (%) on CelebA for different genders.

is not as significant as that on Arcface. We postulate that this is because CIFP leverages sample re-weighting strategy, which is an approximation to causal intervention [27] and has some effects in removing the confounding bias. Nevertheless, our INV-REG further removes the bias and achieves state-of-the-art performance.

RFW. In Table 3, our method achieves the top performance across different races with improved average accuracy and lower standard deviation. In particular, our INV-REG does not require demographic annotations, unlike, *e.g.*, [19]. Note that the accuracy on RFW is near-saturated since it is drawn from the same distribution as the training data.

CelebA. In Figure 5, we show the improvements by adding INV-REG to Arcface (ResNet-100) on each fine-grained attribute. Our model mitigates racial bias, as evidenced by significant improvements in race-related attributes, *e.g.*, “Pale_Skin”, “Blond_Hair” and “Black_Hair”. Furthermore, we effectively eliminate gender bias, *e.g.*, improving gender-related attributes such as “Wearing_Lipstick”, “Heavy_Makeup” and “Mustache”. This is validated by the superior accuracy on Male and Female in Table 4. Besides

	Method	IJB-B	IJB-C
ResNet-50	Arcface [8]	94.33	95.86
	Ours-Arcface	94.44	96.15
	CIFP [43]	94.57	95.80
	Ours-CIFP	94.69	96.15
ResNet-100	Anchorface-Arcface [20]	94.42	96.22
	Anchorface-Curricularface [20]	94.97	96.32
	Arcface [8]	94.20	95.60
	Ours-Arcface	94.94	96.46
	CIFP [43]	95.00	96.47
	Ours-CIFP	95.11	96.58

Table 5: Verification accuracy (%) on IJB-B and IJB-C.

race and gender, our method exhibits robustness across a wide range of attributes, such as accessory, facial hair, hair texture and style. Note that in some attributes our method shows marginal improvements. Possible reasons include: 1) Some attributes are common among a large number of identities in training (*i.e.*, balanced), thus even the baseline captures no bias (*e.g.*, “Smiling”). 2) Some attributes exhibit noisy labels, where faces of different identities are mislabeled as the same one, and we show examples under 3 attributes in Figure 5.

5.3. Conventional Face Recognition

IJB-B and IJB-C. In Table 5, our method demonstrates superiority over the baseline models on both test sets regardless of the backbone choice. In particular, combining our INV-REG with CIFP leads to the highest performance, showcasing the effectiveness of our proposed method.

DigiFace-1M. DigiFace-1M has far greater diversity for each identity and exhibits a domain shift in appearance with the training data. Hence it poses great challenges for model generalization. As shown in Table 6, our method consistently outperforms both baselines by a substantial margin. This is strong proof that our INV-REG captures the causal feature invariant to domain shift and large variations.

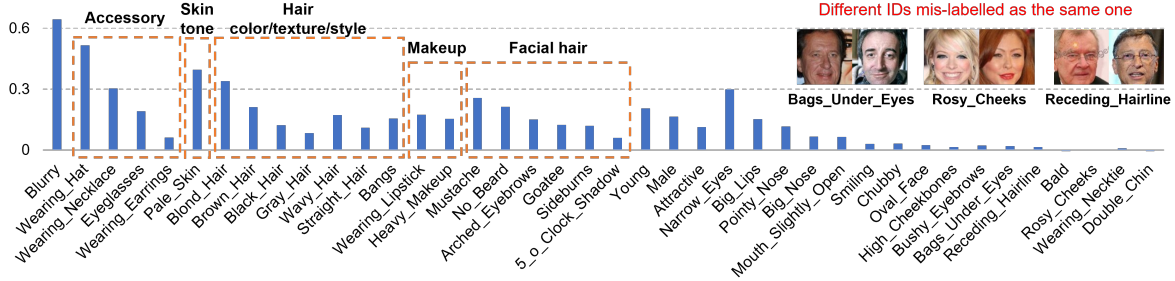


Figure 5: Performance improvement (%) of INV-REG over Arcface on fine-grained attributes of CelebA dataset.

	Method	FPR=1e-5	FPR=1e-4
ResNet-50	Arcface [8]	47.03	62.71
	Ours-Arcface	49.87	64.94
	CIFP [43]	49.36	64.53
	Ours-CIFP	51.19	66.06
ResNet-100	Arcface [8]	48.19	64.32
	Ours-Arcface	51.80	67.57
	CIFP [43]	51.73	67.08
	Ours-CIFP	53.16	68.49

Table 6: Verification accuracy (%) on DigiFace-1M dataset.

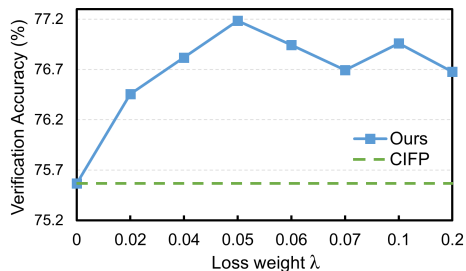


Figure 6: Accuracy (%) on MFR dataset with different λ .

5.4. Ablation Study

Choice of \mathcal{L}_{inv} and Loss Weight λ . We implemented \mathcal{L}_{inv} with REX in Eq. (3) for partition learning and with IRMv1 in Eq. (2) for feature learning, which leads to the best performance. In practice, we find REX is easier to optimize, hence more suitable for the hard task of learning \mathbf{P}^* in Eq. (4), while IRMv1 leads to superior performance in feature learning with its theoretical guarantee [1]. We leave the ablation on \mathcal{L}_{inv} in Appendix due to space constraints. The effect of loss weight λ in \mathcal{L}_{inv} is shown in Figure 6, where the performance is stable in the range of 0.02 to 0.2 with ResNet-50. Note that it is a standard practice in invariant learning to use a small loss weight, hence λ is easy to choose. We adopted the best performing $\lambda = 0.05$ in our experiments.

Learned Partition vs. Ground-truth One. To validate the effectiveness of partition learning, we replaced the learned partitions with the ground-truth race and gender partitions of the training images and followed the same procedure in

Method	AF	CA	SA	EA	Avg	Std
Arcface	74.54	84.43	81.47	53.27	73.43	12.18
Ours [†]	76.55	85.18	83.00	54.61	74.84	12.10
Ours	77.00	85.30	82.88	54.93	75.03	11.99

Table 7: Accuracy (%) on MFR dataset with ground-truth partition (Ours[†]) and learned partition (Ours).

Method	AF	CA	SA	EA	Avg	Std
$K = 2$	79.41	86.53	84.99	57.82	77.19	11.49
$K = 3$	78.97	86.39	84.88	56.67	76.73	11.91
$K = 4$	79.17	86.41	85.11	57.07	76.94	11.79

Table 8: Accuracy (%) on MFR dataset with different K .

#Partitions	AF	CA	SA	EA	Avg	Std
1	78.61	87.15	84.92	56.33	76.75	12.20
2	79.24	86.44	84.99	56.91	76.89	11.85
3	79.41	86.53	84.99	57.82	77.19	11.49
4	79.25	86.38	84.90	56.73	76.82	11.90

Table 9: Accuracy (%) on MFR with different #partitions.

feature learning. The comparison results with ResNet-50 backbone are in Table 7. Learning with the ground-truth partition leads to improved performance compared to Arcface baseline, validating the effectiveness of invariant learning. Furthermore, our learned partition achieves the best performance. We postulate the reason is that the feature-level confounders are complex and elusive, which are not fully captured by the limited attribute annotations on the image level. Our INV-REG discovers diverse feature-level confounders, allowing it to outperform.

#Partition Subsets K . By fixing the optimal setting of $\lambda = 0.05$, we performed ablations on K with ResNet-50 backbone in Table 8. The average accuracy remains stable when we increase K from 2 to 4, with $K = 2$ achieving the overall best performance. Our conjecture is that increasing K makes the optimization more difficult, as the model strives to achieve invariance across all subsets. We will explore an improved optimization strategy in future work.

#Partitions in \mathcal{P} . In Table 9, we observe that discover-

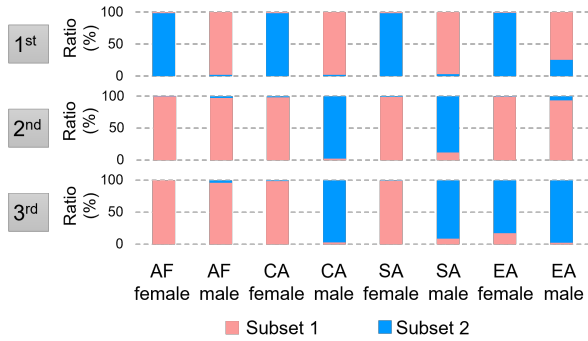


Figure 7: Proportion of the demographic groups in the subsets of each partition (ResNet-50 backbone).

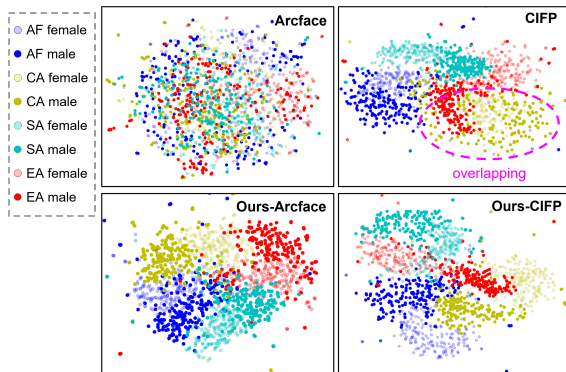


Figure 8: UMAP visualization [24] of features in different methods and demographic groups (ResNet-50 backbone).

ing more partitions generally improves the performance, as the model becomes invariant to more self-annotated confounders. However, using more partitions may require larger #training epochs to fully converge, as evidenced by a slight performance drop with 4 partitions (see Appendix for verification). Hence we adopted 3 partitions to balance the performance and #epochs (all models are trained with 21 epochs).

5.5. Qualitative Analysis

What does the learned partition capture? We visualized the assignment of each demographic group to the 2 subsets of the 3 learned partitions in Figure 7. We observe that each partition captures a demographic attribute. The 1st partition is about gender, with females and males clearly separated. In the 2nd partition, Caucasian and South Asian males are grouped together, possibly because they share common thick facial hair and sharp facial feature (*e.g.*, high cheekbone and prominent nose). The 3rd partition could be about hair texture, *e.g.*, curly or wavy hair is prevalent in subset 1). As mentioned earlier, the feature-level confounders captured by the learned partitions can be elusive. Nevertheless, Table 9 shows that learned partitions are more effective than those based on the ground-truth attributes.

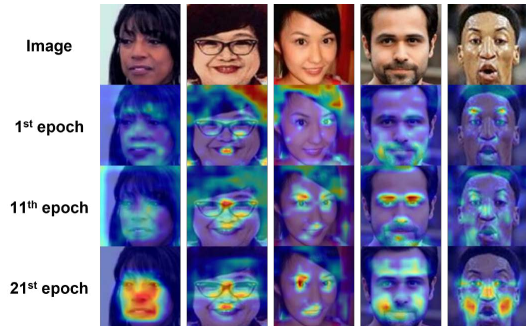


Figure 9: Evolving attention maps during INV-REG learning (ResNet-50 backbone).

What does the invariant feature look like? We visualized the features of 1,600 identities randomly selected from four races and both genders in Figure 8. We observe that compared to Arcface baseline, CIFP has some effects in clustering the features of demographic groups, while our INV-REG has the best clustering quality. We highlight two points: 1) Better clustering indicates improved feature quality by eliminating the confounding demographic effects, such that the demographic features are independent of the rest causal identity features. Hence, images from the same attribute are distributed closely. 2) Bias mitigation is not about removing the demographic attribute features, as they are indeed useful in identity differentiation. The bias lies in the over-dependence of the model on demographic-specific context. We formalize this intuition in Appendix.

How does the attention map evolve in training? Figure 9 visualizes the Grad-CAM [31] in different training epochs. We observe that, as the training progresses, the model learns to capture more causal features (*e.g.*, facial characteristics) while eliminating spurious biases (*e.g.*, hairstyle and facial hair). This further validates our invariant learning.

6. Conclusion

We presented a novel Invariant Regularization (INV-REG) for fair face recognition, which performs causal intervention to remove the confounding effect from the demographic attributes. INV-REG iterates between learning a data partition as a self-annotated confounder, and pursuing invariant feature across the partition subsets to deconfound. Through extensive evaluations on standard benchmarks, we show that INV-REG promotes fairness by improving the accuracy on various minority demographic groups without sacrificing that on the majority ones. In particular, INV-REG is orthogonal to existing methods and can be freely combined with them to achieve improved performance. In future work, we will seek other observational intervention algorithms for improved performance, and pursue an explainable model, *e.g.*, by representation disentanglement [13].

7. Acknowledgements

The authors would like to thank all reviewers and ACs for their constructive suggestions. Part of this research conducted at Nanyang Technological University is supported by the National Research Foundations, Singapore under its AI Singapore Programme (AISG Award No.: AISG2-RP-2021-022).

References

- [1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.
- [2] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023.
- [3] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR, 2019.
- [4] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, and Zhen Lei. Domain balancing: Face recognition on long-tailed domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5671–5679, 2020.
- [5] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arachiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.
- [6] Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pages 2189–2200. PMLR, 2021.
- [7] Jiankang Deng, Jia Guo, Xiang An, Zheng Zhu, and Stefanos Zafeiriou. Masked face recognition challenge: The insight-face track report. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1437–1444, 2021.
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
- [9] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1851–1860, 2017.
- [10] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *IEEE Transactions on Technology and Society*, 1(2):89–103, 2020.
- [11] Sixue Gong, Xiaoming Liu, and Anil K Jain. Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3414–3424, 2021.
- [12] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part III 14*, pages 87–102. Springer, 2016.
- [13] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.
- [14] Isabelle Hupont and Carles Fernández. Demopairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7. IEEE, 2019.
- [15] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020.
- [16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [17] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghui Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.
- [18] Bingyu Liu, Weihong Deng, Yaoyao Zhong, Mei Wang, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Fair loss: Margin-aware reinforcement learning for deep face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10052–10061, 2019.
- [19] Chang Liu, Xiang Yu, Yi-Hsuan Tsai, Masoud Faraki, Ramin Moslemi, Manmohan Chandraker, and Yun Fu. Learning to learn across diverse data biases in deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4072–4082, 2022.
- [20] Jiaheng Liu, Haoyu Qin, Yichao Wu, and Ding Liang. Anchorface: Boosting tar@ far for practical face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1711–1719, 2022.
- [21] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.
- [22] 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.
- [23] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus

- benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.
- [24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [25] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.
- [26] Judea Pearl. *Causality*. Cambridge university press, 2009.
- [27] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.
- [28] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020.
- [29] 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.
- [30] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.
- [31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [32] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.
- [33] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
- [34] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.
- [35] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 692–702, 2019.
- [36] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021.
- [37] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.
- [38] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12241–12248, 2020.
- [39] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016.
- [40] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.
- [41] John D Woodward Jr, Christopher Horn, Julius Gatune, and Aryn Thomas. Biometrics: A look at facial recognition. Technical report, RAND CORP SANTA MONICA CA, 2003.
- [42] Shijie Wu and Xun Gong. Boundaryface: A mining framework with noise label self-correction for face recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 91–106. Springer, 2022.
- [43] Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 578–586, 2021.
- [44] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020.
- [45] Jingzhao Zhang, Aditya Menon, Andreas Veit, Srinadh Bhojanapalli, Sanjiv Kumar, and Suvrit Sra. Coping with label shift via distributionally robust optimisation. *arXiv preprint arXiv:2010.12230*, 2020.
- [46] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017.