ClusterFix: A Cluster-Based Debiasing Approach without Protected-Group Supervision

Giacomo Capitani Federico Bolelli Angelo Porrello
Simone Calderara Elisa Ficarra

Università degli Studi di Modena e Reggio Emilia, Italy
{*name.surname*}@unimore.it

Abstract

The failures of Deep Networks can sometimes be ascribed to biases in the data or algorithmic choices. Existing debiasing approaches exploit prior knowledge to avoid unintended solutions; we acknowledge that, in real-world settings, it could be unfeasible to gather enough prior information to characterize the bias, or it could even raise ethical considerations. We hence propose a novel debiasing approach, termed ClusterFix, which does not require any external hint about the nature of biases. Such an approach alters the standard empirical risk minimization and introduces a per-example weight, encoding how critical and far from the majority an example is. Notably, the weights consider how difficult it is for the model to infer the correct pseudo-label, which is obtained in a self-supervised manner by dividing examples into multiple clusters. Extensive experiments show that the misclassification error incurred in identifying the correct cluster allows for identifying examples prone to bias-related issues. As a result, our approach outperforms existing methods on standard benchmarks for bias removal and fairness.

1. Introduction

Artificial intelligence systems may suffer inductive bias from data (**data bias**), but even algorithmic design choices can expedite wrong decisions during training (**algorithm bias**) [17, 20, 30]. Therefore, enhancing their trustworthiness requires considering more than just performance [3,18,19,25, 45]. More in general, learning systems exhibit the *Principle of Least Effort* [17]: small input changes or different contexts can cause significant deviations in object recognition [2, 35, 43]; minorities or underrepresented groups can suffer high error rates [4, 12, 40].

To address these challenges, there has been widespread interest in debiasing methods that aim to mitigate unintended solutions. Debiasing interventions can occur before the learning procedure (pre-processing), during model training (in-processing), or after training (post-processing) [24]. In particular, in-processing approaches act directly on the algorithm design and effectively mitigate biases. Proposed methods directly debias the model adjusting sample importance [23,36,42] and using adversarial learning [7,26]. Other techniques employ quantitative fairness metrics as regularization [6] and optimization constraints [21]. Although these works address the problem, they require prior knowledge of protected groups, achieved by grouping samples according to their target and bias attribute values, such as gender. Accessing such information is often infeasible due to privacy and ethical constraints. Furthermore, identifying and quantifying bias attributes a priori can pose challenges in complex systems (e.g., organism, climate, and cognition).

So, what if the protected groups are missing during the learning phase?

Our goal is to train a model to mitigate performance disparities among protected groups without exploiting such information during training. We aim to achieve this objective while ensuring satisfactory average performance. These qualities are particularly essential for practical decision-making algorithms. As an example, in healthcare and forensic genetics inaccurate outcomes can pose significant risks to underrepresented communities, claiming for fair and robust solutions [9, 28, 32, 33]. An approximation is needed whenever the system does not access protected groups directly.

Our hypothesis posits that the classification of cluster assignments can leverage shared features between bias-aligned and unbiased samples, presenting an opportunity for an effective mitigation of spurious correlations. Similarly to the works of George [41] and BPA [39], our objective is to address this issue estimating protected groups by clustering.

Existing approaches often encounter a common challenge in the presence of dataset bias: it arises from overfitting noisy and densely populated clusters identified during *Em-*

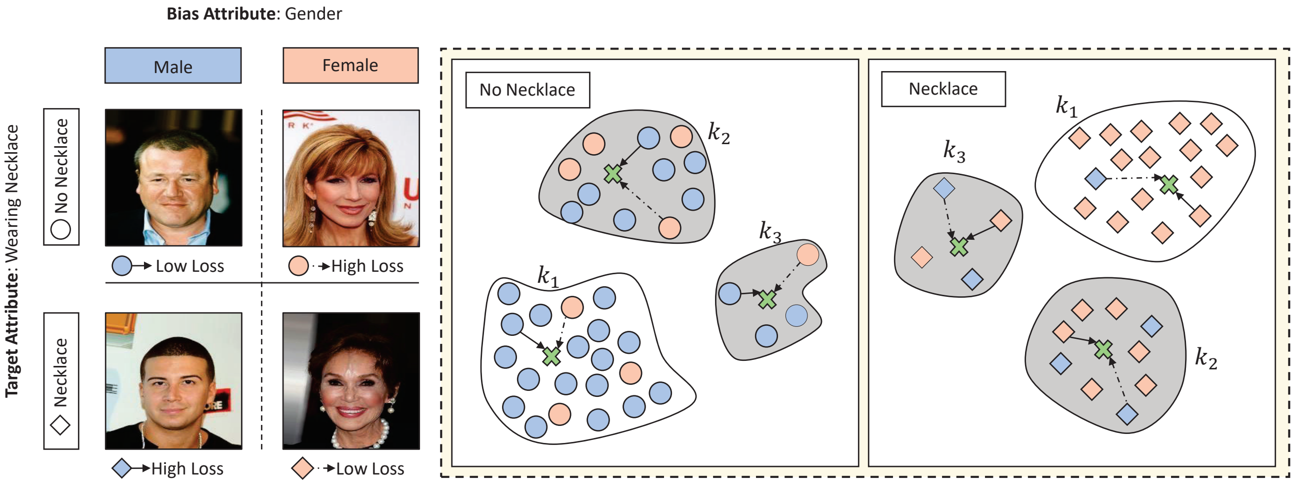
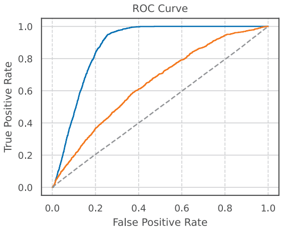


Figure 1. An illustrative scenario depicting dataset partitions biased towards gender attributes. We propose that challenging clusters for classification may leverage significant samples that do not possess the same protected attribute. We emphasize the need to pay closer attention to such distributions in order to address and mitigate model shortcuts.

pirical Risk Minimization (ERM) pre-training, where the focus is solely on minimizing the training loss for the target variable y [39]. Furthermore, these methods rest heavily on the assumption that critical samples can be easily identified through cluster assignments within the ERM feature space. However, real-world scenarios do not guarantee the validity of such an assumption, as critical samples may be dispersed across clusters rather than concentrated in specific ones.

In contrast, our approach takes a different standpoint by acknowledging the dispersed distribution of critical samples within the feature space. Our method involves clustering a self-supervised feature space that operates independently of the label y; in addition to optimizing the target objective, we seek to minimize the cluster classification error. Consequently, we define our weighting mechanism policy by incorporating the auxiliary clustering loss with the target objective. To demonstrate the effectiveness of our approach, we provide a compelling illustration in Fig. 1 with an example.

Identifying Bias. As illustrated in Fig. 1, we consider a dataset partitioned on a target label, such as "Wearing Necklace", where each partition is biased towards a protected attribute like gender. Following the clustering process and assuming these assignments as ground truth, a classifier may encounter challenges within a cluster where the majority of individuals x do not possess the same protected attribute (highlighted in Fig. 1 as critical samples within the grey clusters). To address this challenge, we aim to learn new features that can identify a common per-cluster attribute and achieve satisfactory performance. This means looking for features that are consistent with the original cluster assignment, but different from the protected attribute. We believe that upweighting samples belonging to clusters with high cluster classification loss can aid in identifying data distributions that enhance model robustness and mitigate bias.



Observation. To validate our intuitions, we conducted an experiment on the CelebA dataset (denoted as D) with the target attribute y representing "Wearing Necklace" and the protected attribute a representing gender (female = 0 and male = 1). In addition, we define a group g as $g = (y, a)$. The target label exhibits a strong correlation with female individuals: within the training set D_t, comprising a total of $162,770$ samples, $D_{y=1}$ consists of $18,525$ females and only $1,239$ males. As aforementioned, we defined critical groups as g_0 : $(y = 0, a = 0)$ and g_1 : $(y = 1, a = 1)$. Consequently, the label z define if $x \in \{g_0, g_1\}$ or not. When training an ERM classifier on y, we observed that the accuracy for the group g_1 (male with necklace) was only 2.72%. Subsequently, based on self-supervised features, we obtained cluster assignments for each partition D_y using k-means with $k = 8$. Afterward, we trained an ERM cluster classifier using these assignments. The inset figure presents the ROC curve for partition $D_{a=1}$ (blue) and partition $D_{y=1}$ (orange), which quantifies the correlation between the critical label z and the ERM cluster-objective. Interestingly, we observed a correlation between cluster errors and critical samples in these partitions, indicating the former can be adopted profitably to design an ad-hoc weighting strategy which is the main proposal of our work. In Sec. 3.3, we will formally define how our policy detects critical distributions during the learning phase.

ClusterFix. This paper presents ClusterFix (CFix), a practical in-processing debiasing framework aiming to jointly optimize a classification task and an auxiliary one to generalize over groups without their supervision. Our evaluation demonstrates that the proposed solution outperformed the previous state-of-the-art unsupervised debiasing approaches. Moreover, despite not having group annotations during training, ClusterFix outperformed the supervised approach GDRO [36], even without explicit bias information.

Our contributions can be summarized as follow: (1) We propose an effective debiasing method to mitigate inductive bias in real datasets, which does not rely on prior knowledge on protected groups; (2) We show how cluster assignment classification is a practical auxiliary task that improves worst-case robustness and generalization; (3) Our approach reaches state-of-the-art performances in standard bias-removal benchmarks, even w.r.t. supervised methods.

2. Background

2.1. Related Work

Domain Adaptation and Generalization. In the literature, many works focused on achieving generalized models that can handle worst-case scenarios [48,54]. In this setting, models run into multiple distributions during the learning phase, designed to generalize well on unseen distributions during testing. In robotics, domain randomization has proven to be an effective technique for dealing with real-world situations [44]. Meanwhile, meta-learning aims to acquire the ability to quickly adapt to new conditions by learning how to learn new representations [16,46]. Adversarial training can also be a valuable tool for achieving model robustness and analyzing worst-case scenarios [43, 52]. Adversarial objectives have shown promising generalization abilities and are more aligned with human understanding [37]. On the other hand, they cannot provide a theoretical guarantee of reliable performance [25].

Distributionally Robust Optimization. There has been growing interest in fairness and robustness in recent years. Distributionally Robust Optimization (DRO) proposes to guarantee the best outcomes in cases where there is a change in data distribution by controlling the worst-case performance over subsets [14,42]. However, since worst-case scenarios may be too pessimistic, others investigate group DRO (GDRO) [36], which optimizes worst-case performance over a known set of groups. However, this algorithm requires sensitive-attribute annotation, which is usually unknown in real datasets; we tackle this issue by estimating partitions through clustering.

Fairness without Demographics. Recently, some researchers have explored more challenging scenarios where they have no access to protected-group labels (so-called unsupervised methods). Our framework aims to tackle this issue by replacing sensitive supervision with clustering assignments. Other methods try to achieve sub-population fairness and worst-case generalization without sensitive prior knowledge by estimating different training data distributions. Our work resembles two cluster-based debiasing strategies, George [41] and BPA [39]. Although both methods employ pseudo-labels to partition the dataset into clusters, George has been found vulnerable to outliers overfitting, resulting in suboptimal performance on practical datasets such as CelebA [39].

Deep Clustering and Representation Learning. Our approach leverages clustering to estimate groups and enhance structure-cluster information. In unsupervised settings, clustering serves as a practical training objective for generating improved feature representations. For instance, [5, 11] proposed using cluster assignments as the main training objective to mitigate bias stemming from proxy objectives. In [50], Yan *et al.* enhanced generalization by pre-training a network for clustering in the original feature space and fine-tuning a new model based on cluster assignments. Other approaches have employed deep clustering to jointly learn cluster assignments and representations, preventing feature space collapse [15], or minimizing KL divergence between a centroid-based probability distribution and an auxiliary target distribution [49]. Several clustering methods have recently proposed simultaneously learning fair representations and clustering to conceal sensitive attributes, leveraging protected-group knowledge [1, 10, 26].

2.2. Problem Setup

We are given a set of n datapoints $x_1, ..., x_n \in \mathcal{X}$ associated with a binary label $y_i \in \{0, 1\}$. In addition, each datapoint is associated with a label $c_i \in \{1...C\}$ defined by cluster assignment. We train a deep neural network composed of a feature extractor $\mathcal{F} : X \to R^d$, a binary classifier $\mathcal{T} : R^d \to R^2$, and a set of auxiliary classifiers $\mathcal{C}_y : R^d \to R^C$, one for each task label y. The classification performance of y is evaluated on pre-defined groups G based on the average and worst-group accuracy as in [36]. The group label $g \in \mathcal{G}$ is only used for metric evaluation and is not accessible during optimization. The objective is to achieve good average and worst-group accuracy at test time without training group annotations.

2.3. Relevant Objective Functions

In this section, we describe in detail four training approaches to introduce our work in the next section.

Empirical Risk Minimization. Typically, a model defined by a feature extractor $\mathcal{F}$ and a target classifier $\mathcal{T}$ try to solve the following optimization problem:

$$\arg\min_{\mathcal{F},\mathcal{T}} \frac{1}{N} \sum_{i=1}^{N} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \tag{1}$$

where ℓ represents a loss function, y is the target, and x is the input. In other words, ERM minimizes the average loss across data points. In general, this kind of procedure needs better generalization in some groups during inference.

Group Distributionally Robust Optimization [36]. The Group DRO approach utilizes training group annotations to reduce the maximum group error within the training set. Assuming the availability of group annotations for the training data, the objective function for Group DRO can be expressed as:

$$\arg\min_{\mathcal{F},\mathcal{T}} \max_{g\in\mathcal{G}} \left\{ \frac{1}{N_g} \sum_{i=1}^{N_g} 1(g_i = g)\ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \right\} \tag{2}$$

George [41]. Since g is often not available in real scenarios, George approximates protected groups with cluster assignments. More specifically, this method is organized as follows: (*i*) train a model via ERM, (*ii*) cluster the feature space, (*iii*) train the final model from scratch with clustering information as in Eq. (3). The central empirical hypothesis here is that the feature space of deep neural networks trained via ERM carries information about group labels.

$$\arg\min_{\mathcal{F},\mathcal{T}} \max_{c\in\mathcal{C}} \left\{ \frac{1}{N_c} \sum_{i=1}^{N_c} 1(c_i = c)\ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \right\} \tag{3}$$

BPA [39]. BPA takes advantage of the same hypothesis of George: protected groups are approximated by clustering after ERM pretraining. On the other hand, the optimization process is guided by dynamically reweighting sample importance.

$$\arg\min_{\mathcal{F},\mathcal{T}} \frac{1}{N} \sum_{i=1}^{N} w_{c_i} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \tag{4}$$

where the sample weight w_{c_i} is given by the following equation:

$$w_c = \frac{1}{N_c} \mathbb{E}_{(x,y)\sim P_c} [\ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i)] \tag{5}$$

3. Proposed Method: ClusterFix

3.1. Overview

ClusterFix (CFix) is a two-stage debiasing approach that does not require protected-group supervision. Similarly to previous works, CFix weighs the contribution of each example to the overall classification loss as in Eq. (4); differently, the importance of each example does not depend only on the mismatch between the prediction and ground-truth label y, but also on an additional term that considers how examples cluster in latent space. Briefly, our approach divides into two subsequent steps, as follows.

Cluster Assignment. The first stage regards preparing a pretext task, which will be optimized in the second stage. In particular, the idea is to cluster the latent space of a deep neural network into several groups, thus yielding novel pseudo-labels c from the computed assignments to the clusters. In doing so, CFix leverages self-supervised pre-training (carried out through Barlow Twins [51]), whose latent space is afterward clustered through the k-means algorithm. Such a preliminary stage favorably lowers the risk of biased representation during the downstream task. Indeed, the pseudo-labels produced as such are opaque w.r.t. the target bias-prone task, as they only rely on the self-extracted features.

Debiasing Training. In the second stage, we ask the model to pursue a twofold objective: not only it has to learn the target task embodied by y, but there is a tailored objective that constrains the feature space. In particular, it seeks the model to remain consistent with the original cluster assignments c.Eventually, to preserve minority groups, we weight the importance of each example proportionally to the average *error* of its cluster, where the concept of *error* considers both original and pseudo labels y and c. In this respect, what differentiates CFix from existing approaches is that the clustering membership has an active and direct contribution to the total learning objective.

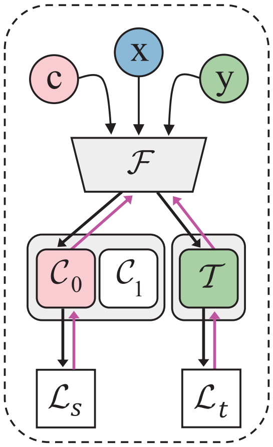


Figure 2. ClusterFix Architecture.

As an example, George [41] as well performs a preliminary clustering step; however, while its authors exploited clustering to compute the classification loss w.r.t. an independent sample grouping (represented by c), we instead use a metric related to c to weight examples.

To provide an intuition, we refer to the observation in Sec. 1 in which a clear dependence arises between an independent clusters' assignment metric (e.g., an ERM pre-trained classifier) and the presence of protected features. For this purpose, this suggests that the empirical risk paid by a wrong cluster assignment can be exploited as a proxy of the importance that should be given to an element in terms of contribution to the task loss on y.

3.2. Step 1: Cluster Assignment

First, the dataset was divided by label, from which features were extracted from a pre-trained model. Next, k-means was applied to obtain cluster assignments c, which were used as categorical labels for the next stage. Specifically, a self-supervised trained model $\mathcal{F} : X \rightarrow R^d$ was used for feature extraction.

3.3. Step 2: Debiased Training

Let $\mathcal{F} : X \rightarrow R^d$, $\mathcal{T} : R^d \rightarrow R^2$, and $\mathcal{C}_y : R^d \rightarrow R^C$ represent the functions of the pre-trained encoder, task classifier, and cluster classifier respectively, Fig. 2. Then, the objective function of the proposed method consists of two parts: weighted-classification loss and cluster-structural loss. In the following, we provide details on these two separate terms and the re-weighting strategy for the proposed model.

Clustering-Structural Loss. Achieving smoothness in the feature space through cluster classification can prevent model shortcuts in y classification and mitigate inductive bias. This method is also valuable for identifying problematic clusters with high average entropy and small size. We hypothesize that there is a correlation between high-entropy clusters and out-of-distribution elements, which can assist the model in generalizing better in worst-case scenarios. This approach's benefits are underpinned by information theory [50], as the structural loss creates an information bottleneck that facilitates the identification of proxy objective bias and improves model generalization. Formally, the clustering-structural loss is defined as follows:

$$\mathcal{L}_s = \sum_{i=1}^{N} \ell(\mathcal{C} \circ \mathcal{F}(x_i), c_i) \tag{6}$$

Task-Weighted Classification Loss. The main objective of the optimization process is to learn how to classify y, which is achieved through task-weighted classification loss. Each sample's classification loss is weighted by a factor w_k, which reflects the significance of the cluster to which the sample belongs. The weight is determined based on the average loss concerning y and the structural loss concerning the clusters. The task-weighted classification loss is defined as follows:

$$\mathcal{L}_t = w_{c_i} \ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) \tag{7}$$

where w_c is:

$$w_c = \frac{1}{N_c} \mathbb{E}_{(x,y) \sim P_c} [\ell(\mathcal{T} \circ \mathcal{F}(x_i), y_i) + \gamma \ell(\mathcal{C}_{y_i} \circ \mathcal{F}(x_i), c_i)] \tag{8}$$

ClusterFix Objective. In summary, the proposed debiasing method's overall objective function can be expressed as the following optimization problem, the trade-off hyperparameter is denoted by γ:

$$\min \mathcal{L}_t + \gamma \mathcal{L}_s \tag{9}$$

4. Experiments

In this section, we present the experiments to evaluate the proposed debiasing method and compare it with several state-of-the-art debiasing techniques. We used the same experimental settings as BPA [39] to ensure a fair comparison among different debiasing methods. Specifically, we trained the proposed method using a ResNet-18 as the backbone architecture with the Adam optimizer, a learning rate of 1×10^{-4}, a batch size of 256 images, and a weight decay rate of 0.01 for 100 epochs. The learning rate was scheduled with cosine annealing. For all experiments, we performed k-means clustering with $K = 8$ and the cluster weight of the c_{th} cluster, ω_c, was updated with a momentum $m = 0.3$ as in [39]. Additional experiments on hyperparameters and datasets can be found in the supplementary material.

4.1. Benchmarks

CelebA. CelebA [27] is a dataset comprising $202,599$ celebrity face images with 40 binary attribute annotations for each image. Moreover, in our experiments, the gender attribute has been used as the bias attribute to evaluate the robustness of the proposed method, as in [39]. We initialized the feature extractor $\mathcal{F}$ parameters by using a self-supervised pre-trained network with Barlow Twins [51]. In particular, for the self-supervised training, we used an output dimension of $1\,024$, a batch size of 256, and an SGD optimizer with a fixed learning rate of 0.6. We set the λ parameter to 0.5.

Following [39], we focused on gender as the fixed bias attribute and excluded 8 out of 40 attributes due to limited samples in the test set. Among the remaining 32 attributes, 26 exhibited a significant correlation with gender, showing a classification accuracy gap of over 5% compared to unbiased accuracy [36]. To explore diverse scenarios, we selected the top 5 attributes with the highest gap and the bottom 5 attributes with the lowest gap, as identified in [39].

Waterbirds. The Waterbirds dataset [36] is designed to evaluate the robustness of deep networks w.r.t. spurious correlations and distribution shifts. It has been created by selecting images of birds from the Caltech-UCSD Birds-200-2011 dataset [47] and overlaying them on backgrounds obtained from the Places dataset [53]. This dataset includes two attributes: the type of bird, which can be either a waterbird or landbird, and the background place, which can be either water or land. The training set comprises $4,791$ samples, 56 out of the $1,113$ waterbird samples have a land background,

Table 1. Unbiased accuracy (%) on CelebA dataset.

	Unsupervised					Supervised
Target	**ERM**	**LfF** [31]	**George** [41]	**BPA** [39]	**Ours**	**GDRO** [36]
Double Chin	64.61 ± 0.82	68.47 ± 0.22	76.23 ± 0.11	82.92 ± 0.54	**85.13 ± 0.30**	83.19 ± 1.11
Pale Skin	71.50 ± 1.60	75.23 ± 0.74	78.22 ± 3.75	90.06 ± 0.75	**91.17 ± 0.04**	90.55 ± 0.84
Chubby	67.42 ± 0.95	71.56 ± 0.52	74.88 ± 1.91	83.88 ± 0.36	**84.16 ± 0.22**	81.90 ± 0.20
Wearing Necklace	55.04 ± 0.59	57.21 ± 0.76	58.79 ± 0.10	68.96 ± 0.12	**68.99 ± 1.19**	62.89 ± 3.69
Wearing Hat	93.53 ± 0.37	94.81 ± 0.15	95.72 ± 0.71	96.80 ± 0.26	**97.88 ± 0.09**	96.84 ± 0.46
Big Lips	60.87 ± 0.58	62.15 ± 0.06	64.99 ± 0.13	**66.50 ± 0.24**	65.40 ± 0.48	63.70 ± 0.44
Bangs	89.04 ± 0.47	89.04 ± 0.50	92.62 ± 0.12	93.94 ± 0.57	**94.67 ± 0.16**	94.45 ± 0.17
Receding Hairline	69.72 ± 0.78	74.58 ± 0.21	78.86 ± 0.40	84.95 ± 0.49	**87.00 ± 0.12**	85.15 ± 1.31
Wavy Hair	73.10 ± 0.56	74.53 ± 0.17	77.39 ± 0.15	**79.89 ± 0.71**	79.42 ± 0.12	79.65 ± 0.63
Brown Hair	78.07 ± 0.87	78.93 ± 1.24	83.07 ± 0.07	83.83 ± 0.66	**85.30 ± 0.47**	84.87 ± 0.07
Average	72.29	74.65	78.07	83.17	**83.91**	82.31

Table 2. Worst-Group accuracy (%) on CelebA dataset.

	Unsupervised					Supervised
Target	**ERM**	**LfF** [31]	**George** [41]	**BPA** [39]	**Ours**	**GDRO** [36]
Double Chin	21.33 ± 2.24	28.24 ± 0.46	50.00 ± 0.41	67.78 ± 0.91	**74.26 ± 3.94**	72.94 ± 1.14
Pale Skin	36.64 ± 3.53	43.26 ± 1.40	62.03 ± 16.50	**88.60 ± 1.48**	87.01 ± 1.46	87.68 ± 2.37
Chubby	24.30 ± 3.73	34.09 ± 0.90	58.01 ± 11.04	72.32 ± 0.93	71.01 ± 1.17	**72.64 ± 1.70**
Wearing Necklace	2.72 ± 0.83	6.67 ± 2.07	13.82 ± 0.41	41.93 ± 2.47	**55.56 ± 0.38**	24.34 ± 7.81
Wearing Hat	85.12 ± 0.31	88.31 ± 0.12	92.93 ± 0.76	94.94 ± 0.19	**96.58 ± 0.63**	94.67 ± 0.41
Big Lips	30.85 ± 0.62	38.54 ± 0.18	44.51 ± 0.83	56.99 ± 3.05	**57.27 ± 0.58**	47.55 ± 1.03
Bangs	76.91 ± 3.27	82.37 ± 0.52	85.90 ± 0.24	92.21 ± 1.24	**93.01 ± 0.36**	92.12 ± 1.03
Receding Hairline	35.69 ± 0.35	45.53 ± 0.55	57.30 ± 0.90	79.11 ± 1.91	**84.15 ± 0.82**	79.12 ± 2.11
Wavy Hair	38.01 ± 0.85	45.24 ± 0.83	53.17 ± 0.43	65.74 ± 1.13	**69.92 ± 0.38**	66.79 ± 1.62
Brown Hair	59.58 ± 2.55	60.68 ± 3.62	73.20 ± 0.88	71.50 ± 0.97	**79.18 ± 0.50**	78.92 ± 1.61
Average	41.11	47.29	59.09	73.11	**76.80**	71.67

while 180 out of $3,678$ landbird samples have a water background. To assess model robustness, the validation and test sets evenly distribute landbirds and waterbirds across land and water backgrounds. We initialized $\mathcal{F}$ with ResNet18 pre-trained on ImageNet and set the λ parameter to 0.01.

Evaluation Protocol. In order to evaluate the proposed method, we calculate the accuracy of every group $g = (y, b)$, defined as a combination of target and bias attribute values. The bias attribute $b \in \{1...B\}$ is an external annotation unused during optimization (e.g., gender, background). We report results in terms of average-group accuracy (**unbiased accuracy**) and **worst-group accuracy** [36, 39]. All reported results are the average of three runs.

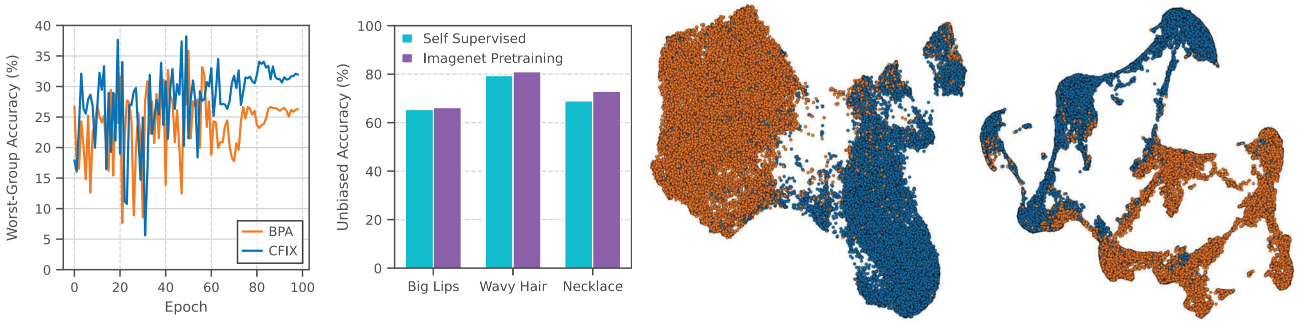
4.2. Comparisons with Other Debiasing Approaches

The proposed method is compared with several state-of-the-art debiasing techniques:

- **Empirical Risk Minimization (ERM)**: our vanilla baseline. It trains a model on a dataset that may be biased, leading to a biased model;
- **Learning from Failure (LfF)** [31]: it trains a debiased classifier using the misclassification information of the biased classifier adopting an adversarial training objective;
- **George** [41]: a debiasing method that approximates bias attributes with cluster assignment and weights the objective function to maximize the worst-case group accuracy;

Table 3. Unbiased and Worst-Group results obtained on the Waterbirds dataset.

		Unbiased Accuracy (%)					Worst-Group Accuracy (%)				
		Unsupervised				Sup.	Unsupervised				Sup.
Target	Bias	ERM	LfF	BPA	Ours	GDRO	ERM	LfF	BPA	Ours	GDRO
Object	Place	84.63	84.57	**87.05**	86.29	88.99	62.39	61.68	71.39	74.03	**80.82**
Place	Object	87.99	85.05	88.44	**92.17**	89.20	73.34	60.00	79.16	**86.61**	85.27



(a) (b) (c) (d)

Figure 3. Robustness evaluation of models trained without critical samples (a) and the effectiveness of CFix backbone pre-training on unbiased accuracy measurement (b). (c) and (d) are UMAP projections visualizing CFix feature space (*Wearing Necklace* = false) at the initial and final epoch. Blue and orange colors represent male and female gender values, respectively.

- **Debiased Representations with Pseudo-Attributes (BPA)** [39]: a debiased representation is learned by introducing cluster-generated pseudo-attributes;
- **Group Distributionally Robust Optimization (GDRO)** [36]: this method optimizes the worst-case performance over a distributionally robust uncertainty set using explicit bias supervision.

Main Results. The results of our evaluation on CelebA are presented in Tab. 1 and Tab. 2. We show how CFix outperforms all competitors across all evaluated scenarios, achieving higher unbiased and worst-group accuracy metrics. Specifically, our method outperforms the bias-supervised approach Group DRO and state-of-the-art BPA in both metrics. Additionally, our optimization process demonstrates greater stability across different runs. The worst-group accuracy improvement is noteworthy, given that other approaches, such as George and Group DRO, prioritize maximizing worst-case group accuracy, which is not the focus of our approach. In more detail, our method achieves an average improvement of $+0.74\%$ in average accuracy and $+3.27\%$ in worst accuracy ($+13.63\%$ for *Wearing Necklace* target) compared to the previous unsupervised state-of-the-art. Additionally, experiments on Waterbirds confirm the effectiveness of our approach even in a controlled environment, as shown in Tab. 3. We observed improvements in the worst-case performance for Object (*bird*) and Place (*backgorund*) classification, with gains of $+2.64\%$ and $+7.45\%$, respectively.

5. Model Analysis

Ratio of the Bias-Aligned Samples. To showcase the resilience of different approaches to distribution shift, we have designed an experiment where we selectively remove bias-aligned samples from the training dataset while keeping the test set unchanged. Our study focuses on the Waterbirds dataset, where we remove minority groups from the training set (all waterbirds in a land background and vice versa). The worst-group accuracy on the test set is reported across epochs in Fig.3a. Our results indicate that CFix outperforms BPA on minority groups not present in the training distribution. This provides evidence that the proposed structural loss enhances the overall worst-case generalization performance.

On the Effects of Pre-Training. The feature extractor $\mathcal{F}$, used for clustering and backbone initialization, is a crucial component in our pipeline. To verify the effectiveness of our approach on CelebA while altering the backbone, ResNet-18 pre-trained on ImageNet has been employed to initialize $\mathcal{F}$. We conducted identical experiments for the best and worst target attributes as specified in Tab. 1. The outcomes illustrated in Fig. 3b indicate minimal changes in the overall performance with a slight improvement when utilizing the latter setting. These results suggest that it is possible to leverage bias-aligned signals without requiring an Empirical

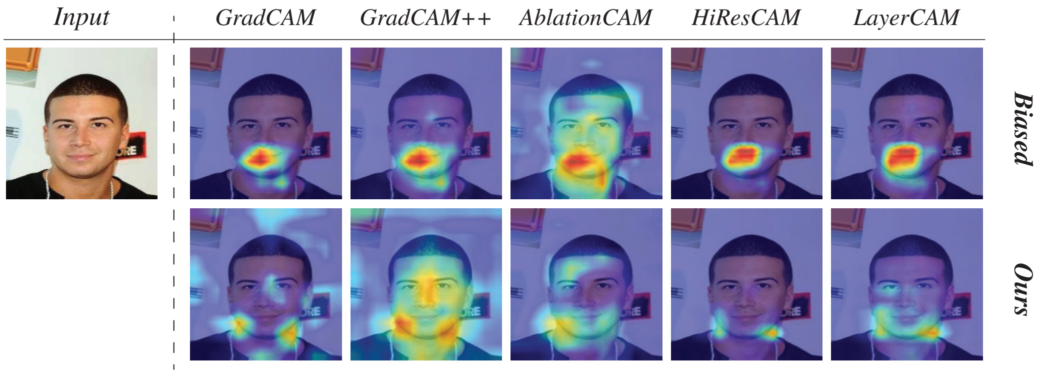


Figure 4. Visualization of the class activation maps generated by Grad-CAM [38], Grad-CAM++ [8], Ablation-CAM [34], HiResCAM [13], and LayerCam [22] for the ERM and the proposed debiased approach targeting the *Wearing Necklace* attribute.

Risk Minimization (ERM) model for each target. Notably, we employed the same self-supervised encoder $\mathcal{F}$ to perform clustering for all CelebA experiments, while others trained a separate ERM model for each target. Refer to the supplementary material for ERM pre-training experiments.

Feature Space Visualization. In Fig. 3c and Fig. 3d, we present visualizations of UMAP [29] projections on the CelebA dataset for the "Wearing Necklace" classification. Specifically, we visualize only negative examples (*Wearing Necklace* = false) to effectively show the feature space at the initial and final epoch using CFix. Our observations suggest that our proposed model successfully achieves a smoother feature space within the same class, mitigating the presence of large clusters caused by shortcuts solutions. Other methods aim to mix the bias attributes in the feature space to improve worst-case generalization. In contrast, our findings advocate that good worst-group and average generalization properties can be achieved even when the bias attributes are separable in the feature space. Therefore, *unlearning* the bias attribute is not always necessary to avoid shortcuts and achieve good model performance.

Model Explainability. The experiments conducted in Fig. 4 explore the interpretability of the proposed debiased approach in contrast to the classical Empirical Risk Minimization (ERM) method. Specifically, we utilize several explainability techniques to visualize the class activation maps of *Wearing Necklace* target on CelebA. For instance, the ERM method emphasizes features not correlated with the target attribute, while CFix prioritizes the regions more indicative of the regions of interest. These results suggest that our method mitigates unintended solutions, providing more meaningful explanations of the decision-making process.

6. Conclusion

Mitigating model shortcuts without directly observing bias attributes is a challenging and relatively unexplored task in achieving bias removal in deep networks. Previous research has attempted to address this problem by modifying the target objective using pseudo-groups identified through ERM pre-training. However, our empirical findings indicate that this approach could be suboptimal in real-world scenarios. Our key contribution is the recognition that empirical cluster error can serve as a proxy for identifying samples likely to be affected by the inductive bias of deep networks. By leveraging this insight, CFix effectively upweights such samples, improving the worst-case and average generalization for protected groups across multiple standard benchmarks. Our study demonstrates that ClusterFix and the insights gained from experimental results offer a robust foundation for advancing worst-case generalization and algorithmic fairness without relying on demographic data.

7. Broader Impact and Limitations

It is crucial to prioritize fairness above performance and always be aware of potential bias, especially in fields such as healthcare and facial recognition. Our novel training approach enhance the worst-case performance for underrepresented groups without relying on their information. Further research should validate whether the proposed approach is valid in various real-world applications beyond the datasets considered in our paper.

Acknowledgments

The work was partially supported by the Italian Ministerial grants PRIN 2022 "B-Fair: Bias-Free Artificial Intelligence methods for automated visual Recognition", CUP E53D23008010006 and "AIDA: explAinable multImodal Deep learning for personAlized oncology", project code 20228MZFAA. The work also received funding from DECIDER, the European Union's Horizon 2020 research and innovation programme under GA No. 965193, and from the Department of Engineering "Enzo Ferrari" of the University of Modena through the FARD-2022.

References

[1] Savitha Sam Abraham. Fairlof: fairness in outlier detection. *Data Science and Engineering*, 6:485–499, 2021.

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[3] Wieland Brendel and Matthias Bethge. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*, 2019.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[6] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 319–328, 2019.

[7] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[8] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[9] Moon S Chen Jr, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. Twenty years post-nih revitalization act: Enhancing minority participation in clinical trials (empact): Laying the groundwork for improving minority clinical trial accrual: Renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

[10] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. *Advances in neural information processing systems*, 30, 2017.

[11] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. 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)*, pages 4069–4082, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[12] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2018.

[13] Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv e-prints*, pages arXiv–2011, 2020.

[14] John Duchi, Tatsunori Hashimoto, and Hongseok Namkoong. Distributionally Robust Losses for Latent Covariate Mixtures. *Operations Research*, 2022.

[15] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-means: Jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138:185–192, 2020.

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[17] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[18] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020.

[19] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[20] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021.

[21] Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *International Conference on Machine Learning*, pages 2879–2890. PMLR, 2019.

[22] Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888, 2021.

[23] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 world wide web conference*, pages 853–862, 2018.

[24] Natasa Krco, Thibault Laugel, Jean-Michel Loubes, and Marcin Detyniecki. When mitigating bias is unfair: A comprehensive study on the impact of bias mitigation algorithms. *arXiv preprint arXiv:2302.07185*, 2023.

[25] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[26] Peizhao Li, Han Zhao, and Hongfu Liu. Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9070–9079, 2020.

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

[28] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine*, 375(7):655–665, 2016.

[29] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[31] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from Failure: Training Debiased Classifier from Biased Classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.

[32] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020.

[33] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 2016.

[34] Harish Guruprasad Ramaswamy et al. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 983–991, 2020.

[35] Amir Rosenfeld, Richard Zemel, and John K Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

[36] Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2020.

[37] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

[38] 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.

[39] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16742–16751, 2022.

[40] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017.

[41] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

[42] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, page 4, 2017.

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[44] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[45] Antonio Torralba and Alexei A Efros. Unbiased Look at Dataset Bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[46] Joaquin Vanschoren. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.

[47] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[48] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[49] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[50] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020.

[51] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

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

[53] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *CoRR*, abs/1610.02055, 2016.

[54] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.