Fairness Meets Cross-Domain Learning: A Benchmark of Models and Metrics

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 3 December 2023, accepted 1 March 2024, date of publication 1 April 2024, date of current version 5 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3383841

RESEARCH ARTICLE

Fairness Meets Cross-Domain Learning: A Benchmark of Models and Metrics

LEONARDO IURADA[1], (Graduate Student Member, IEEE), SILVIA BUCCI[1], TIMOTHY M. HOSPEDALES[2], (Senior Member, IEEE), AND TATIANA TOMMASI[1]

[1]Dipartimento di Automatica ed Informatica (DAUIN), Politecnico di Torino, 10129 Turin, Italy
[2]School of Informatics, The University of Edinburgh, EH8 9YL Edinburgh, U.K.

Corresponding author: Leonardo Iurada (leonardo.iurada@polito.it)

This work was supported in part by Consorzio Interuniversitario CINECA through the Project IsC98_FA-DA under the Italian SuperComputing Resource Allocation–ISCRA Initiative; and in part by the FAIR—Future Artificial Intelligence Research and from European Union (EU) Next-GenerationEU through PNRR—MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3—D.D. 1555 11/10/2022, under Grant PE00000013. The work of Leonardo Iurada was supported by EU Next-GenerationEU [Piano Nazionale di Ripresa e Resilienza (PNRR)] DM 351 on Trustworthy AI. The work of Silvia Bucci was supported by the Blanceflor Foundation. The work of Tatiana Tommasi was supported by the EU Project ELSA—European Lighthouse on Secure and Safe Artificial Intelligence.

ABSTRACT Deep learning-based recognition systems are deployed at scale for real-world applications that inevitably involve our social life. Although of great support when making complex decisions, they might capture spurious data correlations and leverage sensitive attributes (e.g. age, gender, ethnicity). How to factor out this information while maintaining high performance is a problem with several open questions, many of which are shared with those of the domain adaptation and generalization literature which aims at avoiding visual domain biases. In this work, we propose an in-depth study of the relationship between cross-domain learning (CD) and model fairness, by experimentally evaluating 14 CD approaches together with 3 state-of-the-art fairness algorithms on 5 datasets of faces and medical images spanning several demographic groups. We consider attribute classification and landmark detection tasks: the latter is introduced here for the first time in the fairness literature, showing how keypoint localization may be affected by sensitive attribute biases. To assess the analyzed methods, we adopt widely used evaluation metrics while also presenting their limits with a detailed review. Moreover, we propose a new Harmonic Fairness (HF) score that can ease unfairness mitigation model comparisons. Overall, our work shows how CD approaches can outperform state-of-the-art fairness algorithms and defines a framework with dataset and metrics as well as a code suite to pave the way for a more systematic analysis of fairness problems in computer vision (*Code available at: https://github.com/iurada/fairness_crossdomain).*

INDEX TERMS Domain adaptation, domain generalization, fair and trustworthy artificial intelligence, face recognition, landmark detection.

I. INTRODUCTION

Deep neural networks currently constitute the core of several AI systems that support decisions in many socially important tasks such as the hiring process, healthcare diagnosis, and law enforcement. Despite their efficacy, it has become apparent that they can learn to encode subtle biases that disproportionately disadvantage particular sub-populations (*e.g.* based on age, gender, ethnicity, etc.) [1], [2], [3], [4]. The causes of this unfairness are many, from amplifying bias that already exists in the training data [5], to learning spurious correlations [6]. However, the end result is the same: AI systems may exacerbate rather than alleviate social problems of inequality and discrimination. This issue has motivated a growing body of research in fairness interventions [5], [7], [8] — algorithms designed to optimize some notion of fairness jointly with conventional learning objectives. Still, the research in this area is in its infancy and several factors have been overlooked.

This work focuses on the natural alignment of the fair learning problem with the more widely studied cross-domain (CD) learning challenge in computer vision. In the latter area,

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

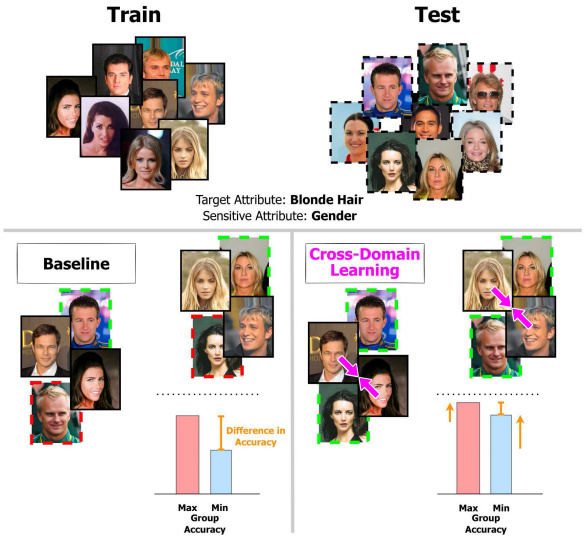

FIGURE 1. In this example the task is predicting if the person in the image has blonde hair or not and we would like to have a model that works equally well for men and women. (Left) We start from a baseline model that exhibits some degree of unfairness as evidenced by the Difference in Accuracy between gender groups. (Right) By exploiting Cross-Domain (CD) learning to reduce the visual domain shift among groups the generalization ability of the model increases, producing an unfairness mitigation effect.

the goal is to produce models agnostic to the specific details of visual domains (*e.g.* camera pose, lighting, image style) to obtain generalization across them. By mapping visual domains to protected subgroups, we can see that the wealth of existing algorithms for promoting domain invariance could potentially benefit fairness (see Figure 1). Thus, **our first contribution** is to present a fairness benchmark for computer vision that spans 5 datasets of face and medical images for classification and landmark detection tasks and compares 14 CD learning approaches alongside 3 state-of-the-art (SOTA) fairness algorithms. We remark that landmark detection on face images of different demographic groups is introduced here for the first time, indeed the bias related to sensitive attributes may affect the whole image annotation but also the precision with which critical keypoints are located.

Another aspect on which there is still a lot of confusion and open debate is about how systems should be evaluated. There are multiple competing notions of fairness and ways to quantify it [7], [9]. Previous studies measured group fairness by accuracy difference between advantaged and disadvantaged subgroups [10]. However, this goal has been criticized in philosophy and ethics literature [11]. Purely minimizing the gap between subgroup performance, may lead to choosing a model with worse accuracy for all subgroups, which is Pareto inefficient [7] and violates the ethical principles of beneficence and non-maleficence [12]. **As our second contribution**, we analyze existing group fairness criteria by highlighting their strengths and weaknesses. The results of our experimental analysis are discussed considering several of those metrics and we also propose a scoring function named Harmonic Fairness that aggregates performance and fairness level to assess the quality of a model and ease the comparison among multiple unfairness mitigation methods.

Finally, our evaluation campaign confirms the effectiveness of CD methods and the relevance of the proposed metric. It highlights how less popular approaches in the CD literature provide a significant advantage for unfairness mitigation on different tasks, systematically outperforming the tailored SOTA approaches. Moreover, it shows that CD models trained to overcome the bias due to one sensitive attribute can be beneficial also to prevent unfairness with respect to a different one. This ability to transfer knowledge provides insights into the robustness of CD approaches for fairness applications. Overall, our work paves the way for a more systematic analysis of fairness problems in computer vision and the related unfairness mitigation methods, providing reliable tools for future evaluations.

II. RELATED WORKS

A. MITIGATING UNFAIRNESS

The concept of fairness is very broad and has been largely discussed in the machine learning literature to support social, economic, and law choices [13], [14], [15], [16]. We focus here on *group fairness* whose aim is to develop decision techniques that are invariant to differences across non-overlapping subsets of data defined by human-sensitive attributes like gender and ethnicity. Several studies have been conducted on face and medical image collections to demonstrate how their biases lead to poor performing recognition models on some minority groups, progressively attracting the attention of the computer vision community [17], [18]. The existing strategies developed to mitigate unfairness have tackled the problem at three main levels depending on when they are applied within the learning process. As data unbalancing is among the main sources of unfairness, some methods act *before training* by collecting ad hoc datasets [19], introducing strategic sampling [5] or developing generative models that mitigate the imbalance through image synthesis [5], [7], [20], [21]. Other techniques have been designed to prevent models from capturing spurious data correlations *during training*, by improving the representation learning procedure. Some approaches quantify these correlations and minimize them by aligning the representations of different demographic groups [22], [23]. Disentanglement-based methods force orthogonality between target classes and sensitive attributes in order to disregard the latter during task learning [24], [25], [26], [27]. A similar goal is obtained by adversarial approaches that include dedicated modules to reduce the discriminability of semantic attributes [5], [28], [29], [30], [31]. Other techniques leverage feature distillation [22], reinforcement [32] and contrastive learning [8]. Very recently, a different family of methods proposed to identify and remove the critical parts of the models causing unfairness [33], [34]. Finally, *post-processing* strategies modify output predictions based on fairness criteria [35].

In terms of tasks, the fairness literature mainly focuses on classification, dealing with both tabular data and images.

Only in the last months, Meta presented a dataset to analyze fairness in classification, detection, and segmentation [36]. Still, we are not aware of previous works discussing fairness for keypoint localization.

B. CROSS-DOMAIN LEARNING

In real-world conditions training and test data often belong to different domains. Cross-domain models are trained to provide good performance on any unseen target domain at test time (*Domain Generalization*), or to adapt the training source knowledge to a specific, different but related target (*Domain Adaptation*).

The techniques proposed to tackle the challenging *Single-Source Domain Generalization* (SSDG) setting extend regularization strategies usually applied in empirical risk minimization to prevent overfitting (*e.g.* label smoothing [37]) and reshape them to face large source-target domain shifts. These include strategic dropout based on gradient observation [38], tailored model selection [39], [40] or data-augmentation to increase data variability [41]. When training samples are drawn from multiple domains, robust models can be obtained via data-augmentation techniques [42], or style-transfer-based approaches [43]. Other popular *Multi-Source Domain Generalization* (MSDG) strategies align the source domain representations through Maximum-Mean Discrepancy (MMD) minimization [44] or adversarial learning [45], [46]. A similar aim is also pursued by multi-task models that combine supervised and self-supervised learning [47], [48]. Meta-learning solutions get prepared for the source-target discrepancy experienced at test time by emulating the same condition with data drawn from the different sources during training [49], [50].

In the *Unsupervised Domain Adaptation* (UDA) setting the target data is available at training time but it is unlabeled. Possible strategies to close the domain gap are based on adversarial learning [51] and feature alignment via MMD [52] or via feature norms matching [53]. Pixel-wise adaptation can also be performed with GAN-based techniques [54], [55]. Clearly, MSDG and UDA share several solutions with slight differences due to the availability of multiple sources in one case, and source and target in the other. Finally, when the target is at least partially labeled, the setting is named *Supervised Domain Adaptation* (SDA) and inherits most of the techniques developed for the more challenging UDA, SSDA, and MSDA. Further constraints are eventually added to prevent overfitting in case of a very limited amount of labeled target data [56], [57], [58], [59].

In terms of tasks, previous works on cross-domain learning broadly cover object classification and detection, as well as semantic segmentation, re-identification and retrieval problems [60]. Still, the task of regression has been significantly less studied [61], [62] and only a few works proposed robust methods for keypoint localization across domains [63], [64], [65], [66].

C. LANDMARK DETECTION

Locating specific points in an image is crucial for applications like face recognition [67], [68], object tracking [69] and pose estimation [70], with practical use in fields such as medicine, sports, and robotics. Keypoints as object corners and edges or facial features like the eyes, nose, and mouth are indicated as landmarks. Older landmark detection methods treat the task as a regression problem, where the goal is to predict continuous pixel coordinates for each landmark [71], [72]. Recent methods have obtained significant gains in accuracy and robustness by modeling the landmark locations through a spatial probability distribution and providing high-resolution 2D heatmaps as output [64], [73]. We consider these heatmap-based strategies in studying the problem of fair landmark detection.

III. FAIRNESS MEETS CROSS-DOMAIN LEARNING

To formalize the problem of fairness we start by defining a data sample as $(\boldsymbol{x}, y, a)$, where $\boldsymbol{x}$ is an image, y is its semantic label and a is a sensitive attribute. In the simplest case, the labels are binary $y \in \{0, 1\}$ (*e.g.* for faces, eye bags yes/no), and the same holds for the attributes $a \in \{0, 1\}$ (*e.g.* male/female or young/old). Given a set of annotated data spanning all the semantic labels and attributes, the goal is to learn a classifier $\hat{y} = f(\boldsymbol{x})$ that correctly predicts the label and achieves certain group fairness criteria with respect to a. These criteria mainly focus on the difference in performance between privileged and disadvantaged data groups associated with distinct attributes (see section IV).

The presented fairness problem shares some common traits with that of cross-domain learning, where source $(\boldsymbol{x}^s, y^s) \sim p^s$ and target $(\boldsymbol{x}^t, y^t) \sim p^t$ data differ on the basis of the distribution from which they are drawn. The information about the distribution is usually summarized by a label indicating the data type: considering one source and one target domain, it holds $d \in \{0, 1\}$ (*e.g.* photos/sketches). By simply switching d with a in the SDA setting we get to the framework described for the fairness problem. As SDA can leverage the whole cross-domain literature, there is a large set of methods that can be applied and evaluated for unfairness mitigation. Some of them have been considered in previous fairness-related publications (*e.g.* discrepancy, adversarial, and disentanglement strategies), but a thorough benchmark is still missing. As discussed in the following, *letting cross-domain learning to meet fairness* may lead to new evaluation strategies and interesting research questions.

IV. FAIRNESS CRITERIA

Evaluating the group fairness of a classification model means assessing its performance on different population subgroups and comparing them. Many criteria have been proposed for this [9], [74]. In the following, we review the most used metrics in computer vision. We start from the basic definitions of True Positive Rate $TPR = TP/(TP + FN)$, False Positive Rate $FPR = FP/(FP + TN)$ and Accuracy $Acc = (TP + TN)/(TP + TN + FP + FN)$. In terms of

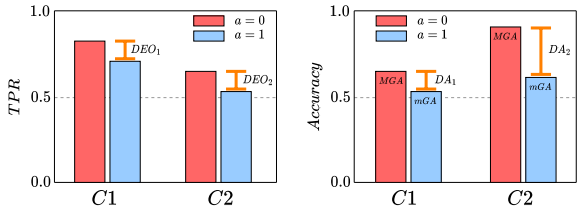

FIGURE 2. Left: $C1$ and $C2$ have the same *DEO* but $C1$ is clearly preferable to $C2$. Right: $C2$ has a higher *mGA* than $C1$, but it also has a higher *DA*. Neither *DEO* nor *mGA* are sufficient for selecting a fair classifier with respect to sensitive attribute *a*.

conditional probabilities for data with two different attributes, it holds

$$TPR_{a=0} = P(\hat{y} = 1|y = 1, a = 0) \tag{1}$$

$$FPR_{a=0} = P(\hat{y} = 1|y = 0, a = 0) \tag{2}$$

and their analogues for $a = 1$. The *Difference in Equal Opportunity (DEO)* measures fairness by

$$|P(\hat{y} = 1|y = 1, a = 0) - P(\hat{y} = 1|y = 1, a = 1)| , \tag{3}$$

so the maximum fairness is obtained for $DEO = 0$ when $TPR_{a=0} = TPR_{a=1}$. The *Difference in Equalized Odds (DEOdds)* measures fairness by

$$\sum_{t\in\{0,1\}} |P(\hat{y} = 1|y = t, a = 0) - P(\hat{y} = 1|y = t, a = 1)|, \tag{4}$$

thus maximum fairness is obtained for $DEOdds = 0$ when both $DEO = 0$ and $FPR_{a=0} = FPR_{a=1}$. In other words, the decision of the classifier should be conditionally independent of the attribute, given the ground truth ($\hat{y} \perp a|y$). Another basic way to consider the variation of the model's output over the subgroups identified by the attributes is via the *Difference in Accuracy (DA)*:

$$|P(\hat{y} = y|a = 0) - P(\hat{y} = y|a = 1)| . \tag{5}$$

All these metrics evaluate the relative behavior of the classifier on data subgroups defined by different attributes but lose track of its absolute performance. This is a critical issue as shown by the practical example in the left part of Figure 2. Although the performance of the two classifiers is different, with $C1$ better than $C2$, they have the same value of *DEO*. Moreover, both *DEO* and *DEOdds* are minimized by a trivial classifier that predicts always $\hat{y} = 1$. In that case, for all the attributes it holds $FN = TN = 0$, so $TPR = FPR = 1$ and $DEO = DEOdds = 0$. Since the accuracy reduces to the Positive Predictive Value ($PPV = TP/(TP + FP)$), also *DA* becomes uninformative.

Recent works have introduced the *Minimum Group Accuracy (mGA)* as fairness criterion: rather than evaluating differences in statistics across groups, it considers the classification accuracy of the worst performing group [7], [30], [75]. The rationale of this metric is that by increasing *mGA* we are certainly improving the overall accuracy. Hence we avoid the suboptimal condition of unnecessarily harming

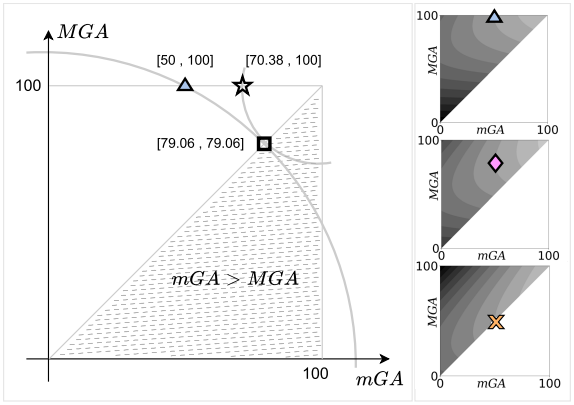

FIGURE 3. Visualization of the [*mGA*, *MGA*] space with exemplar points. The bottom triangular part of the space is unfeasible as by definition *mGA* is lower than *MGA*. The three plots on the right show the HF isolines when starting from different baseline methods indicated by the △, ◇ and × points.

all groups to get a trade-off improvement in fairness measured by *DEO* and *DEOdds*. Still, when the goal is to evaluate whether a certain unfairness mitigation method was able to improve over the reference classifier, *mGA* is not sufficiently informative as exemplified by the right part of Figure 2. Here $a = 0$ is the privileged attribute, thus the one that identifies the best group with the associated *Maximum Group Accuracy (MGA)*. When moving from $C1$ to $C2$, *mGA* increases and so does *MGA*. Although globally the classifier improved, the disadvantaged group suffers even more for unfair treatment with respect to the privileged one as indicated by the increased *DA*.

With these premises, we can state that two unfairness mitigation methods can be reliably compared only by considering at the same time their prediction accuracy and measures of per-group discrepancy. This can be done by looking at several bar plots jointly or at bi-dimensional plots as done in [7]. However, interpreting them and making sense of multiple pieces of information at once is difficult, and defining a single score would facilitate rigorous quantitative evaluations. For this purpose, we can start from the space defined by *mGA* and *MGA*. As shown in Figure 3, the bottom right triangular part of the space is an unfeasible region where $mGA > MGA$. In the top right corner, the point with $[mGA, MGA] = [100, 100]$ indicates the optimal utopia condition. The results of various methods can be collected in this space and ranked on the basis of the $L2$ Distance To the Optimum (*DTO*, [18]) which sounds like a reasonable metric for the score.

Let's focus for instance on the marked points in the figure and consider the biased reference classifier represented by $\triangle = [50, 100]$. We expect a good unfairness mitigation method to keep the top $MGA = 100$ result and improve *mGA* to reduce the discrepancy among groups, thus moving horizontally towards the ideal point. The point $\star = [70.38, 100]$ is a possible result of such an approach. Differently, a method that trades off accuracy for fairness would decrease *MGA* while improving *mGA* to reduce

DA to zero. This behavior is exemplified by the point $\square = [79.06, 79.06]$. It can be noticed that both $\square$ and $\triangle$ share the same Pareto efficiency level approximated by the circumference centered in $[0, 0]$, as done in [7]. Instead, $\star$ shows an efficiency advantage, which is feasible as discussed in [76]. Despite their clear difference, the points $\star$ and $\square$ are equivalent according to *DTO*. Thus, although *DTO* keeps track of both *mGA* and *MGA* it might not be sufficiently informative to benchmark different unfairness mitigation approaches. The presented analysis also highlights the importance of taking as a reference the performance of the baseline to fully understand model comparisons.

V. HARMONIC FAIRNESS

To better deal with the peculiarities of the space defined by *mGA* and *MGA*, we *formalize relative distances* for each method with respect to its biased reference and introduce the *Harmonic Fairness* metric.

A. CLASSIFICATION

We focus on MGA and $DA = MGA - mGA$, using the subscripts b and m to refer respectively to the baseline model and its unfairness-mitigated version. The relative differences are:

$$\Delta DA = DA_b - DA_m \tag{6}$$

$$\Delta MGA = MGA_m - MGA_b \tag{7}$$

with $\Delta DA, \Delta MGA \in \{-100, 100\}$. Both these values will be high for an accurate and fair model. Thus, we combine them in the *Harmonic Fairness* metric defined as:

$$HF = \frac{\Delta DA' \times \Delta MGA'}{\Delta DA' + \Delta MGA'}, \tag{8}$$

where we added an additional shift to the component values to avoid degenerate cases (dividing by 0): $\Delta DA' = \Delta DA + 100$ and $\Delta MGA' = \Delta MGA + 100$. The minimal value $HF = 0$ corresponds to having either $\Delta DA = -100$ or $\Delta MGA = -100$, which can be obtained with a very poorly defined model that reduces the performance (increasing *DA* or decreasing *MGA*) rather than improving over the baseline. An unfairness mitigation model that maintains the same *DA* and *MGA* of the original baseline gets $HF = 50$. Finally, every increase over this value corresponds to models able to symmetrically improve accuracy and fairness.

Getting back to the points $\star$ and $\square$ analyzed before and always considering the $\triangle$ as a baseline, we obtain the meaningful ranking $HF_\star = 54.62 > HF_\square = 51.77$ which matches the expectations given the advantage of the former over the latter. We remark that *HF* takes into proper account the model starting baseline and encourages a decrease in *DA* and an increase in *MGA* with different strengths depending on the baseline position, consequently shaping the space in various ways as shown by the isolines of *HF* in the right part of Figure 3. Of course, the right way to benchmark multiple methods is by setting a fixed baseline model considered as a shared reference for all of them.

B. LANDMARK DETECTION

When dealing with landmark detection every data sample can be defined as $(\boldsymbol{x}, a, \boldsymbol{Y})$, where $\boldsymbol{Y} \in \mathbb{R}^{K\times 2}$ is a set of $\boldsymbol{y}_{1,\dots,K}$ landmark bi-dimensional coordinates. The reference metric for this task is the Normalized Mean Error (*NME*) calculated as:

$$NME(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{K}\sum_{i=1}^{K}\frac{\|\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i\|_2}{D}, \tag{9}$$

where D is a normalization factor, usually chosen as the interocular distance for face images. We indicate with *SDR* the Success Detection Rate calculated as the percentage of images whose NMEs is less than a given threshold. Symmetrically to what was done for classification, we define *Max Group Success* (*MGS*) and *Min Group Success* (*mGS*), respectively as the success rate of the best and worst performing protected groups. We consider also the difference between groups $DS = MGS - mGS$, and to assess the effectiveness of an unfairness mitigation model m over the reference baseline b we calculate:

$$\Delta DS = DS_b - DS_m \tag{10}$$

$$\Delta MGS = MGS_m - MGS_b \tag{11}$$

with $\Delta DS, \Delta MGS \in \{-100, 100\}$. We then combine these values to get the *HF* metric for landmark detection consistent with what is defined for classification in equation (8).

C. RESCALING

To better investigate fine-grained differences among the results of various unfairness mitigation methods we adopt a simple sigmoid rescaling: $\sigma(HF) = \frac{1}{1+\exp\{-HF+50\}}$, with $\sigma(HF) \in \{0, 1\}$. Hence, $\sigma(HF) > 0.5$ will indicate a gain over the reference baseline.

D. DISCUSSION AND EXTENSIONS

To summarize, *HF* is designed to evaluate whether a method is jointly improving in terms of accuracy and fairness. It builds over the worst group accuracy *mGA* (*mGS*) and the best group accuracy *MGA* (*MGS*) by passing through their difference *DA* (*DS*). Still, it is more than a linear combination of the last two terms. If the baseline model is accurate but leads to unfair predictions, then *HF* will rank higher the unfairness mitigation approaches that provide at least the same performance but decrease *DA*. Instead, if the baseline treats each group approximately in the same manner (i.e. low *DA*), then *HF* will rank higher the approaches able to keep at least the same *DA* but that yield better performance.

If needed (*e.g.* in situations where fairness is of utmost importance), the original harmonic mean formulation can be easily adjusted by introducing a weight:

$$HF_w = \frac{\Delta DA' \times \Delta MGA'}{w\Delta DA' + (1-w)\Delta MGA'}, \tag{12}$$

where $w \in [0, 1]$ controls the importance of the two terms. The closer w is to 1, the more $\Delta DA'$ will weigh. Viceversa,

for values of w close to 0, $\Delta MGA'$ will be the prominent term of the equation. For $w = 0.5$ we recover Eq. (8), up to a constant.

Moreover, *HF* can be extended to the case in which the sensitive attribute is multi-class rather than binary by following [8]. Given a sensitive attribute g having G classes, we define the set of all the unique pairs (i, j) of classes within the attribute g as $\mathcal{S} = \{(i, j) \in \{1, \dots, G\} \times \{1, \dots, G\} \mid i \neq j\}$, where "$\times$" represents the cartesian product between sets. We extend *DA* and *MGA* by taking into account the pairs $\forall (i, j) \in \mathcal{S}$ in the following way

$$DA^{ij} = |Acc_i - Acc_j| \tag{13}$$

$$MGA^{ij} = \max_{ij}\{Acc_i, Acc_j\} \,. \tag{14}$$

Similarly to Eq. (6), we derive the quantities

$$\Delta DA^{ij} = (DA_b^{ij} - DA_m^{ij}) \tag{15}$$

$$\Delta MGA^{ij} = (MGA_m^{ij} - MGA_b^{ij}) \,. \tag{16}$$

Thus we can compute *HF* by taking the average over all HF^{ij} values, each of them calculated according to Eq. (8). Mathematically,

$$HF^{ij} = \frac{\Delta DA^{ij'} \times \Delta MGA^{ij'}}{\Delta DA^{ij'} + \Delta MGA^{ij'}} \,, \tag{17}$$

where we applied the same shifts to ΔDA^{ij} and ΔMGA^{ij} of Eq. (8): $\Delta DA^{ij'} = \Delta DA^{ij} + 100$ and $\Delta MGA^{ij'} = \Delta MGA^{ij} + 100$. The resulting formulation for *HF* is

$$HF = \frac{2}{G(G-1)} \sum_{(i,j)\in\mathcal{S}} HF^{ij} \,. \tag{18}$$

If one needs to manage at once multiple sensitive attributes (either binary or multi-class), the fomulation presented above can be repurposed by taking (i, j) from the list of all the unique pairs obtained from the combinations of the classes of each sensitive attribute. Formally, given $g_1, \dots, g_N$ sensitive attributes each having $K_1, \dots, K_N$ classes, respectively, we define $\mathcal{S} = \{(i, j) \in \{1, \dots, K_1\} \times \dots \times \{1, \dots, K_N\} \mid i \neq j\}$. Then, we can apply the very same definition of *HF* given in Eq. (18), where $G = K_1 \cdot \dots \cdot K_N$. As an example, given three sensitive attributes like gender Male(M)/Female(M), age Young(Y)/Old(O) and skin tone Light(L)/Dark(D) we take pairs from the combinations {MYL, MYD, MOL, MOD, FYL, FYD, FOL, FOD}.

VI. BENCHMARK DESCRIPTION

For our analysis, we focus on two tasks that may be affected by fairness issues: attribute recognition and landmark detection. For the former, we consider two binary classification scenarios based on face and medical images. For the latter, we focus on localizing keypoints on face images. To our knowledge, we are the first to study the impact of unfairness on landmark detection and to propose cross-domain learning as a possible solution.

In the following we provide the details of the proposed benchmark that covers 5 datasets, 14 domain adaptive methods (13 for classification and 1 for regression), and 3 SOTA unfairness mitigation approaches.

A. DATASETS

1) CELEBFACES ATTRIBUTE (CELEBA)

[77] comprises 202,599 RGB face images of celebrities, each with 40 binary attribute annotations. We focus on the same subset of 13 reliable target attributes considered in [7] and [78]. We select *male* and *young* as protected attributes, and adopt the same setting of [7], based on the official train/val/test splits.

2) COVID-19 CHEST X-RAY

Reference [79] is composed of 719 images of chest x-ray coming from different online sources showing scans of patients affected by pulmonary diseases. Each image has a structured label describing many attributes of the patient. We focus on the *finding* attribute as target, considering the COVID-19 pathology, while *gender* is selected as sensitive attribute. We split the dataset into 80/20% training/test sets, using 20% of the training split for validation.

3) FITZPATRICK17K

Reference [80] is a collection of 16,577 clinical images depicting 114 skin conditions from two dermatology atlases. The images are annotated with the six Fitzpatrick skin type labels, that describe the skin phenotype's sun reactivity. The dataset is widely used in algorithmic fairness research [18]. We classify whether the dermatological condition in each picture is either *benign/non-neoplastic* or *malignant* and we use *skin tone* as the protected attribute, keeping only the examples belonging to *skin type I* (light) and *skin type VI* (dark) of the Fitzpatrick scale. We split the dataset into 80/20% training/test sets, using 20% of the training split for validation.

4) UTKFACE

Reference [81] consists of over 20k RGB face images characterized by great variability in terms of pose, facial expression, illumination, etc., and present age, gender, and race annotations. We focus on landmark localization (68 points) considering the values *white* and *black* of the label *race* as protected groups for the experiments related to *skin tone*. Moreover, we define the *young* and *old* groups by collecting respectively samples with the value of label *age* in 0-10 and 40-50 years old. Training/test division is 80/20% with 20% of the training split used for validation.

5) FAIRFACE

Reference [19] is a dataset designed to assess model fairness in face attribute recognition. It comprises a vast collection of 108,501 facial images, representing individuals from seven race groups: *White*, *Black*, *East Asian*, *Southeast*

Asian, *Middle Eastern* and *Latino*. The dataset provides also annotations for *gender* and *age* groups. We adopt the same setup of [82]: *gender* as sensitive attribute, multi-class *race* as target.

B. REFERENCE METHODS

1) BASELINES

For our classification experiments we follow the fairness literature [7], [83] adopting as baseline ResNet50 with standard cross-entropy minimization objective, pre-trained on ImageNet. For landmark detection we follow [64] and [84] and consider ResNet18 pre-trained on ImageNet with a dedicated head composed of deconvolutional layers. It is optimized with an $L2$ loss to reduce the discrepancy between the predicted probability distribution of the location of each landmark and the ground truth.

2) FAIRNESS REFERENCES

We consider three SOTA unfairness mitigation methods. GroupDRO [83] minimizes the worst-case training loss over a set of pre-defined groups. FSCL [8] re-designs supervised contrastive learning to ensure fairness by paying attention to the choice of the negative samples and to the distribution of the anchors between data groups. Finally, g-SMOTE [7] is a generative approach that reduces unfairness by synthesizing new samples of the most disadvantaged group. All of them focus on classification problems while we are not aware of works dedicated to unfairness mitigation on landmark detection.

3) CROSS-DOMAIN METHODS

We investigate methods from four main families. The *Regularization-based approaches* include all the techniques designed to prevent overfitting with a consequent boost in the model generalization ability. LSR [37] encourages the model to avoid overconfidence by smoothing data annotation. SWAD [39] searches for flat minima. RSC [38] is based on a refined drop-out. L2D [41] includes a module trained to synthesize new images with a style distribution complementary to that of the training data. The models based on *Adversarial training* encode domain-invariant representations by preventing the network from recognizing the domains. In DANN [51] the gradient computed by a domain discriminator is inverted while learning the data representation. CDANN [45] improves over DANN by matching the conditional data distributions across domains rather than the marginal distributions. Finally, SagNets [85] introduces dedicated data randomizations to disentangle style from class encodings. *Feature alignment* models involve training objectives that minimize domain distance measures. AFN [53] measures domain shift by comparing the feature norms of two domains and adapts them to a common large value. MMD [88] minimizes the homonym metric to reduce the domain discrepancy. Lastly, Fish [86] proposes to align the domain distributions by maximizing the inner product between their gradients. *Self-Supervised Learning*-based techniques exploit auxiliary self-supervised tasks to let the network focus on semantic-relevant features. RelRot [87] predicts the relative orientation between a reference image (anchor) and the rotated counterpart as an auxiliary task. Here we also consider a variant that we name RelRotAlign to encourage the domain alignment using as anchor a sample with the same target attribute but from a different protected group. SelfReg [46], exploited contrastive losses to regularize the model and guide it to learn domain-invariant representations.

4) LANDMARK DETECTION

The community has dedicated less attention to domain adaptive approaches for keypoint detection. For our analysis, we consider the recent RegDA [64] that was developed to target human pose estimation and introduced an adversarial regressor based on the Kullback-Leibler divergence between domains to narrow their gap. We also extend DANN [51] and AFN [53] to this task.

VII. EXPERIMENTS

In this section we present the main results of our experiments. The code at the basis of our evaluation is in Pytorch and covers all the methods in the benchmark, providing maximal transparency and fostering reproducibility. We organized the code as a suite that can also easily welcome other methods for future benchmark extensions. Unless stated otherwise, for all the experiments we adopted the same validation protocol described in [7]. Further information about the implementation details can be found in the appendix.

A. CLASSIFICATION RESULTS

For the binary classification tasks, we present the tables with different horizontal sections that group the cross-domain methods by family. The bottom part of the tables contains the SOTA fairness approaches. The columns show the evaluation metrics with the aim of providing an overview of model accuracy as well as fairness criteria already discussed in section IV. Specifically, for *DTO* we use the relative formulation $\Delta DTO = DTO_b - DTO_m$: as the baseline is fixed and shared by all the methods, ΔDTO ranks the methods exactly as *DTO* but makes the tables easier to read.

The results on CelebA are presented in Table 1 and focus on the two most challenging attributes: *EyeBags* and *Chubby*. Out of the whole set of 13 attributes, they are the ones with the lowest *Acc* and the highest *DA*.

In terms of overall accuracy *Acc*, most of the approaches provide a small improvement or are equivalent to the baseline. A similar trend can be observed by looking at the per-group accuracy with a few notable exceptions among the CD methods that present more visible gains over the baseline as well as over the SOTA competitors: AFN shows the best performance in terms of *mGA*, followed by DANN on Eyebags and SagNets on Chubby. Still, the *mGA* metric cannot be considered alone: for instance for EyeBags,

TABLE 1. Results obtained on face images when the task is to recognize *EyeBags* (left) and *Chubby* (right) with gender as sensitive attribute. Every number represents the average over three runs. Bold indicates the best results, underline the second best.

		CelebA - EyeBags (*gender*)								CelebA - Chubby (*gender*)							
		Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	$\sigma(\mathbf{HF})$↑	Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	$\sigma(\mathbf{HF})$↑
	Baseline [78]	81.46	88.59	70.15	18.44	20.75	39.10	0.00	0.500 ± 0.000	94.95	98.54	89.24	9.30	27.55	31.61	0.00	0.500 ± 0.000
Reg.	**LSR** [37]	82.27	89.03	71.53	17.50	22.67	41.93	1.45	0.584 ± 0.034	94.96	98.58	89.22	9.36	27.37	31.25	-0.01	0.499 ± 0.014
Reg.	**SWAD** [39]	80.25	87.24	69.15	18.09	16.97	44.67	-1.42	0.444 ± 0.184	94.20	98.47	87.43	11.04	43.85	51.67	-1.80	0.393 ± 0.122
Reg.	**RSC** [38]	82.52	89.07	72.13	16.94	15.44	32.19	2.02	0.621 ± 0.020	94.89	98.40	89.33	9.06	29.39	33.98	0.06	0.506 ± 0.013
Reg.	**L2D** [25]	81.70	88.31	71.20	17.11	18.31	34.77	0.88	0.563 ± 0.056	95.07	98.64	89.41	9.23	28.30	32.16	0.18	0.510 ± 0.006
Adv.	**DANN [51]**	**83.82**	**90.28**	73.56	16.72	33.71	48.89	3.79	0.700 ± 0.043	95.07	98.56	89.53	9.03	23.49	26.81	0.29	0.518 ± 0.011
Adv.	**CDANN [45]**	81.71	87.75	72.11	15.64	8.11	22.91	1.50	0.610 ± 0.078	94.89	98.53	89.12	9.41	36.05	40.94	-0.12	0.492 ± 0.020
Adv.	**SagNets** [85]	83.48	90.17	72.85	17.32	28.57	45.82	3.09	0.660 ± 0.075	**95.18**	**98.66**	89.65	9.01	26.67	30.35	0.42	0.526 ± 0.009
Feat.	**AFN** [53]	83.59	89.34	**74.47**	14.87	**3.55**	**12.05**	**4.29**	$\mathbf{0.744} \pm 0.039$	95.16	98.51	**89.85**	**8.66**	22.58	25.17	**0.60**	$\mathbf{0.536} \pm 0.021$
Feat.	**MMD** [44]	83.53	89.94	73.35	16.59	21.01	36.57	3.47	0.689 ± 0.020	95.11	98.55	89.64	8.90	24.98	28.69	0.40	0.525 ± 0.030
Feat.	**Fish** [86]	82.45	89.43	71.38	18.05	29.27	51.81	1.45	0.575 ± 0.038	94.95	98.57	89.19	9.39	33.42	37.84	-0.04	0.496 ± 0.030
Self.	**RelRot** [87]	82.88	89.55	72.29	17.26	22.98	42.66	2.35	0.629 ± 0.063	95.01	98.49	89.48	9.02	26.99	31.23	0.23	0.507 ± 0.027
Self.	**RelRotAlign**	74.87	76.57	72.29	**4.28**	33.69	42.31	-4.33	0.414 ± 0.113	94.67	98.10	89.23	8.87	**2.04**	**3.57**	-0.08	0.493 ± 0.007
Self.	**SelfReg** [46]	83.59	90.14	73.20	16.94	21.78	36.72	3.40	0.679 ± 0.058	95.07	98.61	89.46	9.15	32.12	36.09	0.23	0.513 ± 0.015
	GroupDRO [83]	83.08	89.25	73.28	15.98	16.30	30.82	3.16	0.681 ± 0.067	94.88	98.43	89.22	9.21	30.14	34.09	-0.03	0.499 ± 0.017
	g-SMOTE [7]	82.00	88.94	72.38	16.56	28.11	46.63	2.21	0.632 ± 0.071	94.61	98.51	88.41	10.10	27.25	33.77	-0.83	0.448 ± 0.026
	FSCL [8]	82.89	89.56	72.29	17.27	34.51	46.02	2.35	0.619 ± 0.192	95.00	98.48	89.48	9.00	15.53	19.42	0.23	0.518 ± 0.010

TABLE 2. Results obtained on CelebA considering the whole set of 13 reliable attributes as target, while gender is the sensitive attribute. The numbers represent the average over 13 experiments, each repeated 3 times. Bold indicates the best results, underline the second best.

		CelebA - 13 Attributes (*gender*)							
		Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	$\sigma(\mathbf{HF})$↑
	Baseline [78]	91.90	93.71	89.60	4.12	15.05	18.47	0.00	0.500 ± 0.000
Reg.	**LSR** [37]	91.97	93.76	89.72	4.04	14.85	18.27	0.13	0.507 ± 0.010
Reg.	**SWAD** [39]	91.45	93.50	88.92	4.58	17.23	22.20	-0.69	0.458 ± 0.034
Reg.	**RSC** [38]	91.89	93.57	89.80	3.77	14.44	17.81	0.10	0.513 ± 0.007
Reg.	**L2D** [25]	91.91	93.74	89.66	4.08	14.91	18.03	0.07	0.504 ± 0.008
Adv.	**DANN [51]**	**92.15**	93.92	89.96	3.95	16.02	19.08	0.42	0.523 ± 0.001
Adv.	**CDANN [45]**	92.04	93.73	89.92	3.80	13.46	16.25	0.28	0.520 ± 0.016
Adv.	**SagNets** [85]	92.08	93.87	89.86	4.00	17.14	20.58	0.31	0.516 ± 0.014
Feat.	**AFN** [53]	92.14	93.85	**90.02**	3.83	13.24	15.97	**0.43**	$\mathbf{0.527} \pm 0.008$
Feat.	**MMD** [44]	92.09	93.83	89.93	3.90	15.24	18.35	0.34	0.521 ± 0.008
Feat.	**Fish** [86]	91.85	93.69	89.58	4.11	15.55	19.02	-0.03	0.499 ± 0.000
Self.	**RelRot** [87]	92.10	93.86	89.90	3.95	15.47	18.97	0.33	0.519 ± 0.006
Self.	**RelRotAlign**	91.14	92.39	89.70	**2.69**	**8.29**	**10.50**	-0.65	0.503 ± 0.014
Self.	**SelfReg** [46]	91.98	93.76	89.75	4.01	15.50	18.18	0.15	0.510 ± 0.011
	GroupDRO [83]	91.97	93.78	89.77	4.01	14.05	17.10	0.18	0.511 ± 0.005
	g-SMOTE [7]	92.12	**93.94**	89.88	4.05	14.68	18.47	0.36	0.517 ± 0.007
	FSCL [8]	91.58	93.50	89.14	4.36	13.22	15.78	-0.50	0.473 ± 0.002

RelRot and RelRotAlign have the same *mGA* but they differ significantly in terms of the other accuracy scores *Acc* and *MGA*. Moreover, the low value of *DA* for RelRotAlign on EyeBags comes together with an improvement in *mGA* over the baseline (72.29>70.15) and with a significant loss in *MGA* (76.57<81.46). Thus, even looking only at *DA* may be misleading. The same holds for *DEO* and *DEOdds* as indicated by their low values for RelRotAlign on Chubby. These last two metrics reward the *leveling down* behavior already criticized in [7], by largely relying on *DA* without considering the decrease in *MGA* and *mGA* with respect to the baseline.

Finally, ΔDTO and $\sigma(HF)$ have a similar role here in providing indications of the models' trustworthiness by summarizing the previous metrics. They assign RelRotAlign a very low rank with respect to the other competitors and agree to identify AFN as the best method.

For completeness, we also present in Table 2 the results on the set of 13 attributes already used in [7] and [78]. According to [78], these attributes are the most reliable out of the whole set of 40 CelebA attributes as they can be labeled objectively, without being ambiguous for a human. AFN confirms itself as the best method as it is able to increase both *mGA* and *MGA*, while decreasing *DA*. The second best is DANN confirming the effectiveness of adversarial techniques to deal with the fairness problem [5], [89]. Considering the high baseline accuracy, the improvements of the different methods appear relatively small but they are consistent with the results presented in the supplementary material of [7].

The results on COVID-19 Chest X-Ray and Fizpatrick17k are presented in Table 3. On the first dataset, according to both ΔDTO and $\sigma(HF)$, RSC is the top method and RelRotAlign is the second best, while AFN ranks third. Even in this case, it is clear that referring only to *mGA* may not be sufficient to differentiate among the methods as many of them share the exact same value for this metric. Specifically, this happens for RSC and RelRotAlign: by observing *DA* one should rank the second as better than the first. This possible confusion underlines again the benefit of summary metrics.

The results on Fizpatrick17 lead to similar conclusions, with RelRot and LSR presenting the best results. DANN, which was among the top methods for CelebA, now ranks sixth among all the CD approaches and still shows results comparable with the best SOTA fairness approach.

Overall, the exact CD family that best suits each classification task may vary (feature alignment and adversarial training methods for faces, regularization-based and self-supervised approaches for medical images), but the results confirm the effectiveness of cross-domain learning for unfairness mitigation and the relevance of our study.

B. MULTI-CLASS SENSITIVE ATTRIBUTES AND TARGET

The data in Fizpatrick17 are annotated with six skin types that range from light to dark. This provides the possibility to investigate the effectiveness of CD approaches even in a multi-class attribute setting by comparing the extended versions of *DEO*, *DEOdds*, and $\sigma(HF)$. We focus on the methods with the most promising performance in the previous analysis and present the results in Table 4. They confirm that cross-domain models outperform SOTA fairness-dedicated approaches with increased accuracy for all skin tones.

By exploiting the FairFace dataset we can instead study the multi-class target case. Specifically, we considered gender as binary sensitive attribute and multi-class race as target,

TABLE 3. Results on medical images for covid recognition with gender as sensitive attribute (left) and for benign/malignant skin lesion recognition with skin tone as sensitive attribute. Every number represents the average over three runs. **Bold** indicates the best results, underline the second best.

		COVID-19 Chest X-Ray (*gender*)								Fitzpatrick17k (*skin tone*)							
		Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	σ(**HF**)↑	Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	σ(**HF**)↑
Baseline [78]		73.79	78.62	65.41	13.22	26.44	35.18	0.00	0.500 ± 0.000	90.22	94.45	87.26	7.19	11.51	13.29	0.00	0.500 ± 0.000
Reg.	**LSR** [37]	68.96	72.83	62.26	10.57	22.41	26.83	-5.84	0.307 ± 0.144	92.89	93.66	92.25	1.41	19.65	20.53	3.88	0.766 ± 0.020
Reg.	**SWAD** [39]	73.10	79.35	62.26	17.09	19.49	25.48	-2.36	0.348 ± 0.278	91.59	93.20	90.44	2.76	10.31	11.29	2.16	0.678 ± 0.054
Reg.	**RSC** [38]	**77.93**	**80.43**	**73.58**	6.85	8.62	11.32	**7.79**	**0.880** ± 0.035	91.72	92.70	90.89	1.81	20.63	21.07	2.22	0.698 ± 0.046
Reg.	**L2D** [25]	70.34	71.74	67.92	3.82	12.07	21.14	-2.09	0.574 ± 0.049	92.56	**94.77**	90.91	3.86	16.52	17.59	3.41	0.709 ± 0.021
Adv.	**DANN [51]**	71.03	71.74	69.81	**1.93**	8.62	**10.83**	-0.69	0.666 ± 0.068	92.04	92.98	91.32	1.66	11.38	13.44	2.73	0.720 ± 0.046
Adv.	**CDANN** [45]	70.34	72.83	66.04	6.79	17.24	22.88	-2.83	0.492 ± 0.116	91.85	92.95	90.99	1.95	7.40	9.15	2.46	0.704 ± 0.045
Adv.	**SagNets** [85]	74.48	79.35	66.04	13.31	15.52	27.04	0.92	0.540 ± 0.042	91.46	92.70	90.58	2.12	15.10	15.86	1.98	0.683 ± 0.033
Feat.	**AFN** [53]	76.55	79.35	71.70	7.65	13.79	16.98	5.63	0.821 ± 0.045	91.95	93.60	90.72	2.87	15.42	16.74	2.62	0.694 ± 0.033
Feat.	**MMD** [44]	76.55	**80.43**	69.81	10.62	13.79	24.58	4.69	0.741 ± 0.113	91.60	93.30	90.43	2.87	8.84	10.28	2.21	0.679 ± 0.018
Feat.	**Fish** [86]	75.86	78.26	71.70	6.56	17.24	24.10	4.98	0.812 ± 0.072	92.24	93.52	91.39	2.12	17.14	18.79	3.12	0.723 ± 0.078
Self.	**RelRot** [87]	74.48	76.09	71.70	4.39	10.34	16.47	3.62	0.793 ± 0.093	**93.35**	94.47	**92.61**	1.87	**2.01**	**2.81**	**4.67**	**0.786** ± 0.017
Self.	**RelRotAlign**	75.86	77.17	**73.58**	3.59	12.07	22.36	5.75	0.857 ± 0.073	91.65	92.81	90.70	2.11	12.02	13.36	2.14	0.687 ± 0.039
Self.	**SelfReg** [46]	73.79	77.17	67.92	9.25	17.24	18.71	1.29	0.642 ± 0.067	92.69	94.61	91.27	3.34	8.34	10.14	3.64	0.727 ± 0.029
	GroupDRO [83]	70.34	73.91	64.15	9.76	22.41	29.77	-3.67	0.403 ± 0.095	91.98	92.64	91.41	**1.24**	10.19	11.60	2.58	0.722 ± 0.037
	g-SMOTE [7]	73.14	77.97	64.11	13.86	25.60	34.49	-1.45	0.420 ± 0.043	90.92	93.19	88.50	4.69	12.26	14.27	0.53	0.569 ± 0.066
	FSCL [8]	60.02	63.04	54.86	8.18	**1.72**	15.20	-17.68	0.051 ± 0.035	90.67	93.60	88.48	5.13	12.42	13.85	0.72	0.566 ± 0.176

TABLE 4. Results on medical images for benign/malignant skin lesion recognition with skin tone as multi-class sensitive attribute. Every number represents the average over three runs. **Bold** indicates the best results, underline the second best.

	Fitzpatrick17k - Benign/Non-neoplastic vs. Malignant (*skin tone multi-class*)								
	Acc_1↑	Acc_2↑	Acc_3↑	Acc_4↑	Acc_5↑	Acc_6↑	DEO↓	DEOdds↓	σ(**HF**)↑
Baseline	90.82	91.45	93.90	91.41	95.31	93.09	14.63	106.78	0.500 ± 0.000
DANN	95.21	95.82	95.67	97.71	96.26	95.76	15.30	109.87	0.722 ± 0.034
AFN	92.41	92.26	93.93	96.60	97.18	96.36	12.04	92.27	0.620 ± 0.013
RelRot	**97.82**	**98.20**	**98.58**	**98.80**	**99.44**	**98.48**	**10.71**	**70.70**	**0.827** ± 0.006
GroupDRO	94.93	95.96	96.57	97.52	98.43	98.39	11.78	77.38	0.742 ± 0.008
g-SMOTE	93.09	92.17	93.08	90.57	96.27	92.31	16.25	104.37	0.512 ± 0.031
FSCL	91.86	92.99	93.62	91.87	94.44	95.71	13.30	86.33	0.552 ± 0.027

TABLE 5. Results on FairFace for the recognition of seven race groups with gender as binary sensitive attribute. Every number represents the average over three runs. **Bold** indicates the best results, underline the second best.

	FairFace - Race (*gender*)							
	Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	σ(**HF**)↑
Baseline	71.92	72.88	71.21	1.68	25.43	25.43	0.00	0.500 ± 0.000
DANN	72.32	72.89	**71.83**	1.06	**6.02**	**6.02**	0.45	**0.539** ± 0.014
AFN	**72.63**	**73.12**	71.81	1.31	14.35	14.35	**0.60**	0.537 ± 0.012
GroupDRO	71.74	71.84	71.65	**0.19**	26.47	26.47	-0.41	0.525 ± 0.023
g-SMOTE	71.03	71.79	70.58	1.21	20.92	20.92	0.21	0.523 ± 0.011
FSCL	71.24	71.64	71.04	0.60	18.61	18.61	-0.98	0.488 ± 0.031

TABLE 6. Landmark detection results. SDR is evaluated using 8% NME as threshold. Results averaged over three runs.

	UTKFace Landmark (*skin tone*)					
	SDR↑	MGS↑	mGS↑	DS↓	ΔDTO↓	σ(**HF**)↑
Baseline	82.08	83.90	77.93	5.97	0.00	0.500 ± 0.000
AFN [53]	92.95	93.80	90.99	**2.81**	16.38	0.961 ± 0.047
DANN [51]	89.62	90.66	86.17	4.49	10.63	0.883 ± 0.017
RegDA [64]	**96.05**	**97.05**	**93.77**	3.28	**20.43**	**0.979** ± 0.011
	UTKFace Landmark (*age*)					
Baseline	74.50	78.39	70.71	7.67	0.00	0.500 ± 0.000
AFN [53]	94.50	95.04	**93.97**	**1.06**	28.59	**0.997** ± 0.001
DANN [51]	81.03	85.22	76.83	8.39	8.92	0.809 ± 0.091
RegDA [64]	**94.62**	**95.51**	93.74	1.77	**28.70**	0.996 ± 0.001

following the same setup [82]. The obtained results are shown in Table 5 with the EO metrics calculated according to the definition in [90]. Even in this setting the CD methods confirm their effectiveness.

C. LANDMARK DETECTION RESULTS

The performance of a model which locates keypoints on facial components may be affected by a change in *skin tone* and *age*, resulting in a less precise prediction in case of high melanin pigmentation or wrinkles. To investigate the presence of a bias related to these demographics we run experiments on the UTK Face dataset and we verify the effectiveness of correction strategies based on cross-domain learning by considering RegDA together with AFN and DANN, as they have shown successful results in classification on face images. The training procedure follows the one presented in [64], with validation protocol in line with that of [7]. We assess the performance of the methods by considering both $\sigma(HF)$ and ΔDTO obtained from *SDR* calculated with a standard 8% *NME* threshold [73].

Table 6 shows how the baseline reference has an unfair behavior with more than 5% difference in group accuracy (*DS*). All the cross-domain methods provide an advantage: in particular, RegDA ranks higher or equal to AFN, and they are both better than DANN. The latter shows a large improvement in *MGS* and *mGS* when the sensitive attribute is age, but the group discrepancy appears worse than the baseline.

By reducing the *NME* threshold the evaluation becomes progressively more demanding until the extreme of considering a predicted point as successful only if it perfectly overlaps with the ground truth. The curves in Figure 4 show that even moving toward this condition most of the cross-domain methods maintain their advantage over the baseline confirming their effectiveness. The difference between RegDA and AFN becomes more evident at lower threshold values. In that regime, *HF* (as well as $\sigma(HF)$) and ΔDTO show different trends for RegDA with the first discouraging the use of this approach when the sensitive attribute is age.

Although no previous publication proposed an unfairness mitigation approach for landmark detection, GroupDRO might sound general enough to be applied also in this setting. This approach dynamically adjusts loss weights during optimization to prioritize the poorest-performing protected group. However, our investigation revealed that, even after a comprehensive hyperparameter search, the loss of the worst group decreases extremely slowly in landmark detection, and

TABLE 7. Model Transferability analysis on the classification task. All the relative metrics are calculated with respect to the baseline results in the first row. The bottom part of the table presents oracle results, i.e. a reference upper bound for the top part of the table. Every number is obtained as the average over three runs. Bold indicates the best results, underline the second best. Note that these fonts are used only when the results improve over the baseline, thus they do not appear in most of the columns of the right part of the table.

CelebA - EyeBags																
	Male/Female ⟶ *Young/Old*								*Young/Old* ⟶ *Male/Female*							
	Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	σ(**HF**)↑	Acc.↑	MGA↑	mGA↑	DA↓	DEO↓	DEOdds↓	ΔDTO↑	σ(**HF**)↑
Baseline [5]	81.06	83.20	74.05	9.16	11.54	24.69	0.00	0.500 ± 0.000	82.55	89.17	72.04	17.13	19.72	41.95	0.00	0.500 ± 0.000
AFN [53]	81.31	83.30	74.92	**8.38**	9.06	21.42	0.78	0.554 ± 0.014	81.35	88.64	69.77	18.87	23.71	50.51	-2.31	0.361 ± 0.068
DANN [51]	**83.48**	**85.83**	**76.16**	9.67	18.28	31.74	**3.18**	**0.626** ± 0.039	**82.64**	**89.50**	71.75	17.75	21.83	42.73	-0.15	0.481 ± 0.021
GroupDRO [83]	82.52	84.90	75.68	9.22	15.48	29.15	2.29	0.599 ± 0.055	82.10	89.36	70.58	18.79	32.38	56.68	-1.30	0.408 ± 0.098
g-SMOTE [7]	80.16	82.58	72.62	9.96	8.89	22.18	-1.54	0.415 ± 0.168	82.01	88.77	71.38	17.34	15.66	36.10	-0.76	0.458 ± 0.132
FSCL [8]	80.45	84.65	69.35	15.30	**8.64**	**14.38**	-3.37	0.235 ± 0.097	82.32	89.30	71.25	18.05	**11.73**	**28.31**	-0.69	0.450 ± 0.029
	Young/Old ⟶ *Young/Old* (**Oracle**)								*Male/Female* ⟶ *Male/Female* (**Oracle**)							
AFN [53]	81.70	83.81	75.04	8.77	10.92	23.43	1.16	0.561 ± 0.013	83.59	89.34	74.47	14.87	3.55	12.05	2.32	0.646 ± 0.032
DANN [51]	83.00	84.90	77.01	7.89	7.42	16.74	3.41	0.676 ± 0.040	83.82	90.28	73.56	16.72	33.71	48.89	1.81	0.594 ± 0.049
GroupDRO [83]	82.23	83.70	75.67	8.03	16.99	35.92	1.63	0.598 ± 0.049	83.08	89.25	73.28	15.98	16.30	30.82	1.18	0.575 ± 0.075
g-SMOTE [7]	81.21	81.99	73.71	8.28	6.63	17.97	-0.95	0.477 ± 0.072	82.00	88.94	72.38	16.56	28.11	46.63	0.23	0.521 ± 0.013
FSCL [8]	80.50	84.66	69.39	15.27	8.72	14.44	-3.33	0.237 ± 0.098	82.89	89.56	72.29	17.27	34.51	46.02	0.37	0.515 ± 0.023

TABLE 8. Model transferability analysis on the landmark detection task. All the relative metrics are calculated with respect to the baseline results in the first row. The bottom part of the table presents oracle results, i.e. a reference upper bound for the top part of the table. Every number is obtained as the average over three runs. Bold indicates the best results, underline the second best.

UTKFace - Landmark Detection												
	Skin Tone ⟶ *Age*						*Age* ⟶ *Skin Tone*					
	SDR↑	MGS↑	mGS↑	DS↓	ΔDTO↑	σ(**HF**) ↑	SDR ↑	MGS↑	mGS↑	DS↓	ΔDTO↑	σ(**HF**) ↑
Baseline	88.30	90.14	85.13	5.01	0.00	0.500 ± 0.000	82.35	86.67	77.97	8.70	0.00	0.500 ± 0.000
AFN [53]	**91.46**	**93.08**	**89.84**	3.23	**5.55**	**0.760** ± 0.116	88.03	91.67	84.33	7.34	8.00	0.828 ± 0.152
DANN [51]	79.55	80.91	78.07	**2.85**	-11.23	0.126 ± 0.074	79.83	80.84	78.82	**2.02**	-2.81	0.501 ± 0.033
RegDA [64]	89.64	91.82	87.64	4.18	3.02	0.650 ± 0.092	**90.55**	**94.59**	**86.44**	8.15	**11.15**	**0.883** ± 0.040
	Age ⟶ *Age* (**Oracle**)						*Skin Tone* ⟶ *Skin Tone* (**Oracle**)					
AFN [53]	91.22	92.03	90.49	1.55	5.43	0.792 ± 0.009	91.32	91.94	90.68	1.27	13.43	0.960 ± 0.020
DANN [51]	80.10	80.77	79.45	1.32	-10.30	0.160 ± 0.099	83.20	86.03	78.85	7.18	0.40	0.551 ± 0.060
RegDA [64]	91.28	91.61	90.61	1.00	5.25	0.796 ± 0.043	92.69	94.78	89.88	4.90	14.36	0.950 ± 0.014

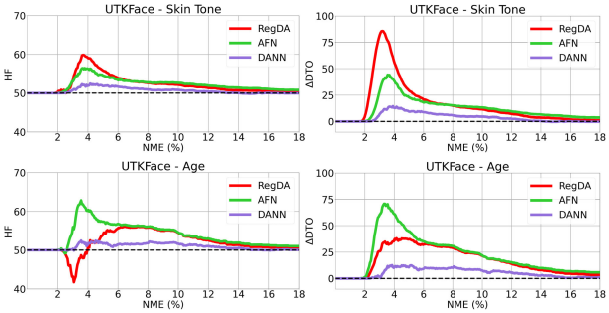

FIGURE 4. Landmark detection results. Comparison among the cross-domain methods and the reference baseline in terms of *HF* and *DTO* when changing the NME threshold used for *SDR*.

the method keeps focusing on it which ultimately makes it unable to obtain an improvement neither on the best group nor overall. The result is a high Normalized Mean Error (NME) achieved by GroupDRO during training and a consequent 0% Success Detection Rate (SDR) on the test set. Notably, applying looser thresholds did not improve the situation, suggesting that the logic employed by GroupDRO may not be well-suited for landmark detection tasks.

VIII. MODEL TRANSFERABILITY

Considering the effort needed to train novel models, it is always desirable to exploit existing ones for new tasks. For unfairness mitigation approaches, what is learned by reducing the bias over some protected groups might be helpful also for other demographics. We study this aspect on the CelebA dataset considering *EyeBags* as the target attribute with *Male* and *Young* as sensitive attributes.

We train and validate a classifier to recognize whether eye bags are present while learning to disregard gender-specific features through a cross-domain approach. Then, we test the obtained model by assessing how the eye bags prediction performance differs among age groups. We analyze the CD methods AFN and DANN, reporting also the results of the SOTA unfairness mitigation strategies. The top left part of Table 7 shows the effect on age groups of the approaches trained to be gender agnostic while focusing on the semantic features relevant to identifying the target *EyeBags* attribute. The results exceed those of the baseline with a particular advantage of DANN over GroupDRO, indicating that the knowledge acquired with cross-domain learning is transferrable. The bottom left part of the table presents the performance of *oracle* methods trained and validated with the aim of mitigating age bias. They represent an upper bound and allow to better appreciate the surprisingly competitive results of transferred cross-domain models. We note how the SOTA unfairness mitigation models obtain low results even in this oracle setting.

The right part of Table 7 shows the results obtained when inverting the roles for the sensitive attributes (*Young/Old* → *Male/Female*). Now a model trained to be agnostic to age is tested to evaluate whether it helps in mitigating unfairness with respect to *gender*. From the results we observe that

the transferability is not symmetric as none of the methods improve over the baseline.

The transferability results look instead always very promising on landmark detection: a model trained to be fair on *skin tone* is effective also in reducing the performance gap among different *age* groups and vice-versa as shown in Table 8, which may be explained by a moderate correlation between skin surface-based cues (such as pigmentation or shadows cast by the 3D shape of the face) with perceived age.

Overall the possibility to reuse fair models on different sensitive attributes connects with the ability of the models to capture knowledge shared across them and generalize at deployment time to new social conditions with different ethical constraints. We find it an interesting aspect that gives rise to new research questions and deserves more attention in the future.

IX. CONCLUSION

In this paper we proposed an extensive study on the problem of fairness in computer vision by presenting a new benchmark to assess the performance of cross-domain learning approaches for unfairness mitigation. Our work covers several demographics and goes beyond classification by introducing landmark detection in fairness research. Passing through a review of the existing criteria used to evaluate unfairness mitigation methods, we proposed Harmonic Fairness to summarize several relevant metrics and ease comparisons among various approaches. Finally, we shed light on the reusability of models created to be robust among specific protected groups by underlying their effectiveness when facing different sensitive attributes. We dedicated particular attention to the reproducibility of our study by releasing the code of the implemented methods in a suite that can be easily extended for future analysis.

Although our focus is mainly on group fairness and other definitions are possible, we believe that our work provides several tools to broaden the study of fairness-related issues and solutions in AI.

APPENDIX A
IMPLEMENTATION DETAILS

A. CLASSIFICATION

For all the experiments we follow [7] in terms of base architecture, training details, and validation protocol. In particular, all the methods are built upon the ImageNet pre-trained ResNet-50 [91] backbone optimized with Adam ($lr = 10^{-4}$, batch size 64). As data augmentation, we use a center crop to 128×128 and RandAugment with $N = 3$ and $M = 15$. The validation is done every 500 iterations and the best model is selected based on the best *mGA* computed on the validation set. Note that for g-SMOTE [7] we used the GAN inversion model provided in [92], pre-trained on CelebA: the official GAN code and weights used in [7] have not been released by the authors. Although there may be some debate around the use of generative approaches that are not tailored specifically to the medical task at hand, we decided to incorporate g-SMOTE into both the experiments on the COVID-19 Chest X-Ray and Fitzpatrick17k datasets for completeness.

We perform an extensive hyper-parameters search to find the best models for every approach considered in our benchmark. In particular, we apply the Random Search [93] algorithm followed by a refinement stage in the following hyper-parameter intervals:

LSR [37]: ε is the coefficient used to smooth the ground truth labels such that $y_k^{LS} = y_k(1-\varepsilon) + \varepsilon/K$, where K indicates the number of classes. Used range: $\{\varepsilon \in [0.1, 0.5]\}$;

SWAD [39]: r is the tolerance rate used on the validation loss function when searching the interval in which the model's parameters have to be sampled and averaged. We didn't tune the optimum patience (N_e) and the overfit patience (N_s) since the overfitting behavior could be observed already after the very first validation. Used range: $\{r \in [0.1, 1.3]\}$;

RSC [38]: f indicates the dropping percentage to mute the spatial feature maps, b indicates the percentage of the batch on which RSC is applied. Used range: $\{f \in [10, 80], b \in [10, 80]\}$;

L2D [25]: α_1 weights the contribution of the supervised contrastive loss function and α_2 weights the negative log-likelihood between the latent vectors of the source image x and the generated one x^+ in the final objective function. Used range: $\{\alpha_1 \in [0.1, 3.0], \alpha_2 \in [0.1, 3.0]\}$;

DANN [51], CDANN [45]: λ is the hyper-parameter that weights the reverse gradient during the backpropagation step, γ controls the penalty assigned to the norm computed on the gradients of the domain discriminator. Used range: $\{\lambda \in [0.01, 1.00], \gamma \in [0.01, 0.50]\}$;

SagNets [85]: λ weights the adversarial loss function. Used range: $\{\lambda \in [0, 2]\}$;

AFN [53]: λ trades off the feature-norm penalty and the supervised cross-entropy loss, R is the value at which the norms of the extracted features are forced to converge to. Used range: $\{\lambda \in [0.01, 0.10], R \in [5, 100]\}$;

MMD [44]: γ weights the MMD loss term in the final objective. Used range: $\{\gamma \in [0.1, 1.0]\}$;

Fish [86]: η weights the gradient inner product. Used range: $\{\eta \in [0.01, 0.10]\}$;

RelRot, RelRotAlign [87]: α weights the importance of the self-supervised loss function in the total objective. Used range: $\{\alpha \in [0.1, 1.0]\}$;

SelfReg [46]: $\lambda_{feature}$ and λ_{logit} control, respectively, the in-batch dissimilarity losses applied to the intermediate features and the logits from the classifier. Used range: $\{\lambda_{feature} \in [0.1, 1.0], \lambda_{logit} \in [0.1, 1.0]\}$;

GroupDRO [83]: C is a model capacity, η is the step size to update the weights and balance worst/best performing groups. Used range: $\{\eta \in [0.001, 0.05], C \in [1, 10]\}$;

g-SMOTE [7]: m in the number of nearest neighbors considered, k is the number of random points chosen among the m and λ is the probability of selecting a batch from the original dataset during training. Used range: $\{m \in [2, 10], k \in [2, m], \lambda \in [0.1, 1.0]\}$.

B. LANDMARK DETECTION

Throughout our experiments, we adopt the architecture and training procedures outlined in [64]. To ensure consistency, we also use the validation protocol proposed in [7]. Our approach employs an ImageNet pre-trained ResNet-18 [91] backbone, followed by an upsampling head consisting of three 2D transposed convolutions with a dimension of 200 and a kernel size of 4. This head performs heatmap regression to determine the position of each landmark, resulting in an output tensor $\hat{\mathbf{Y}} \in \mathbb{R}^{200\times200\times68}$. Our network is optimized using stochastic gradient descent (SGD) with a learning rate of 0.1, momentum of 0.9, weight decay of 1e-4, and a batch size of 32 for 35000 iterations. We incorporate a multi-step learning rate decay with a decay factor of 0.1, using iteration 22500 and 30000 as milestones. To apply several augmentation sequentially we use the TorchLM library.[1] The augmentations are: random rotation (with angles ranging from -180 to 180 degrees), random horizontal flip (with a probability of 0.5), random shear (with x and y rescale factors of 0.6 and 1.3, respectively), color jitter (with brightness, contrast, and saturation set to 0.24, 0.25, and 0.25, respectively) and Gaussian blur (with a kernel size of 5 and $\sigma = (0.1, 0.8)$). We validate every 500 iterations and select the best model based on the highest *mGS* score on the validation set.

We conduct an exhaustive search for optimal hyperparameters for all the approaches included in our benchmark. Specifically, we employ the Random Search algorithm [93], followed by a refinement stage, within the hyperparameter intervals as specified below:

AFN [53]: λ trades off the feature-norm penalty and the supervised cross-entropy loss, R is the value at which the norms of the extracted features are forced to converge to. Used range: $\{\lambda \in [1e-6, 0.10], R \in [5, 100]\}$;

DANN [51]: λ is the hyper-parameter that weights the reverse gradient during the backpropagation step, γ controls the penalty assigned to the norm computed on the gradients of the domain discriminator. Used range: $\{\lambda \in [1e-6, 1.00], \gamma \in [0.01, 0.50]\}$;

RegDA [64]: margin trades off between the KL divergence loss and the Regression Disparity loss. t is a modifier of the magnitude of the Regression Disparity loss. Used range: $\{margin \in [1.0, 10.0], t \in [0.01, 1.0]\}$.

REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. FAccT*, 2018, pp. 77–91.

[2] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociol. Methods Res.*, vol. 50, no. 1, pp. 3–44, Feb. 2021.

[3] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Med.*, vol. 27, no. 12, pp. 2176–2182, Dec. 2021.

[4] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

[5] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8916–8925.

[6] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[7] D. Zietlow, M. Lohaus, G. Balakrishnan, M. Kleindessner, F. Locatello, B. Schölkopf, and C. Russell, "Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10400–10411.

[8] S. Park, J. Lee, P. Lee, S. Hwang, D. Kim, and H. Byun, "Fair contrastive learning for facial attribute classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10379–10388.

[9] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. IEEE/ACM Int. Workshop Softw. Fairness (FairWare)*, May 2018, pp. 1–7.

[10] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. NeurIPS*, 2016, pp. 1–12.

[11] A. Mason, "Egalitarianism and the levelling down objection," *Analysis*, vol. 61, no. 3, pp. 246–254, Jul. 2001.

[12] T. L. Beauchamp, "Methods and principles in biomedical ethics," *J. Med. Ethics*, vol. 29, no. 5, pp. 269–274, Oct. 2003.

[13] T. Brennan and W. L. Oliver, "The emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions," *Criminol. Public Policy*, vol. 12, no. 3, pp. 551–562, Aug. 2013.

[14] E. van den Broek, A. Sergeeva, and M. Huysman, "Hiring algorithms: An ethnography of fairness in practice," in *Proc. ICIS*, 2019, pp. 1–9.

[15] R. Berk, "Accuracy and fairness for juvenile justice risk assessments," *J. Empirical Legal Stud.*, vol. 16, no. 1, pp. 175–194, Feb. 2019.

[16] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY, USA: Crown, 2016.

[17] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein, "Discrimination in the age of algorithms," *J. Legal Anal.*, vol. 10, pp. 113–174, Dec. 2018.

[18] Y. Zong, Y. Yang, and T. Hospedales, "MEDFAIR: Benchmarking fairness for medical imaging," in *Proc. ICLR*, 2023, pp. 1–10.

[19] K. Kärkkäinen and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1547–1557.

[20] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 2261–2268.

[21] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness GAN: Generating datasets with fairness properties using a generative adversarial network," *IBM J. Res. Develop.*, vol. 63, no. 4/5, pp. 3:1–3:9, Jul. 2019.

[22] S. Jung, D. Lee, T. Park, and T. Moon, "Fair feature distillation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12110–12119.

[23] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 692–702.

[24] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *Proc. FAccT*, 2018, pp. 119–133.

[25] J. Lee, E. Kim, J. Lee, J. Lee, and J. Choo, "Learning debiased representation via disentangled feature augmentation," in *Proc. NeurIPS*, 2021, pp. 25123–25133.

[26] E. Tartaglione, C. A. Barbano, and M. Grangetto, "EnD: Entangling and disentangling deep representations for bias correction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13503–13512.

[27] M. H. Sarhan, N. Navab, A. Eslami, and S. Albarqouni, "Fairness by learning orthogonal disentangled representations," in *Proc. ECCV*, 2020, pp. 746–761.

[1] https://github.com/DefTruth/torchlm

[28] P. Dhar, J. Gleason, A. Roy, C. D. Castillo, and R. Chellappa, "PASS: Protected attribute suppression system for mitigating bias in face recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15067–15076.
[29] T. Kehrenberg, M. Bartlett, O. Thomas, and N. Quadrianto, "Null-sampling for interpretable and fair representations," in *Proc. ECCV*, 2020, pp. 565–580.
[30] E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth, "Minimax group fairness: Algorithms and experiments," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2021, pp. 66–76.
[31] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *Proc. ECCV*, 2020, pp. 330–347.
[32] M. Wang and W. Deng, "Mitigating bias in face recognition using skewness-aware reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9319–9328.
[33] Y. Zhang, S. Gao, and H. Huang, "Recover fair deep classification models via altering pre-trained structure," in *Proc. ECCV*, 2022, pp. 481–498.
[34] Y. Savani, C. White, and N. S. Govindarajulu, "Intra-processing methods for debiasing neural networks," in *Proc. NeurIPS*, 2020, pp. 2798–2810.
[35] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proc. AAAI*, 2019, pp. 247–254.
[36] L. Gustafson, C. Rolland, N. Ravi, Q. Duval, A. Adcock, C.-Y. Fu, M. Hall, and C. Ross, "FACET: Fairness in computer vision evaluation benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20370–20382.
[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[38] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Proc. ECCV*, 2020, pp. 124–140.
[39] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "SWAD: Domain generalization by seeking flat minima," in *Proc. NeurIPS*, 2021, pp. 22405–22418.
[40] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *Proc. UAI*, 2018.
[41] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 814–823.
[42] M. Xu, J. Zhang, B. Ni, T. Li, C. Wang, Q. Tian, and W. Zhang, "Adversarial domain adaptation with domain mixup," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 4, pp. 6502–6509.
[43] F. C. Borlino, A. D'Innocente, and T. Tommasi, "Rethinking domain generalization baselines," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9227–9233.
[44] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5400–5409.
[45] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, "Deep domain generalization via conditional invariant adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 647–663.
[46] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, "SelfReg: Self-supervised contrastive regularization for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9599–9608.
[47] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2224–2233.
[48] S. Bucci, A. D'Innocente, Y. Liao, F. M. Carlucci, B. Caputo, and T. Tommasi, "Self-supervised learning across domains," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5516–5528, Apr. 2021.
[49] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. AAAI*, 2018, pp. 3490–3497.
[50] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. Hospedales, "Episodic training for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1446–1455.
[51] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Apr. 2016.
[52] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 97–105.
[53] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1426–1435.
[54] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[55] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[56] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8049–8057.
[57] B. Li, Y. Wang, S. Zhang, D. Li, K. Keutzer, T. Darrell, and H. Zhao, "Learning invariant representations and risks for semi-supervised domain adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1104–1113.
[58] L. Yang, Y. Wang, M. Gao, A. Shrivastava, K. Q. Weinberger, W.-L. Chao, and S.-N. Lim, "Deep co-training with task decomposition for semi-supervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 8886–8896.
[59] T. Kim and C. Kim, "Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation," in *Proc. ECCV*, 2020, pp. 591–607.
[60] G. Csurka, T. M. Hospedales, M. Salzmann, and T. Tommasi, *Visual Domain Adaptation in the Deep Learning Era*, 1st ed. San Rafael, CA, USA: Morgan & Claypool, 2022.
[61] X. Chen, S. Wang, J. Wang, and M. Long, "Representation subspace distance for domain adaptation regression," in *Proc. ICML*, 2021, pp. 1749–1759.
[62] J. Wu, J. He, S. Wang, K. Guan, and E. Ainsworth, "Distribution-informed neural networks for domain adaptation regression," in *Proc. NeurIPS*, 2022, pp. 10040–10054.
[63] L. O. Vasconcelos, M. Mancini, D. Boscaini, S. R. Bulò, B. Caputo, and E. Ricci, "Shape consistent 2D keypoint estimation under domain shift," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8037–8044.
[64] J. Jiang, Y. Ji, X. Wang, Y. Liu, J. Wang, and M. Long, "Regressive domain adaptation for unsupervised keypoint detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6776–6785.
[65] T. Ohkawa, Y.-J. Li, Q. Fu, R. Furuta, K. M. Kitani, and Y. Sato, "Domain adaptive hand keypoint and pixel localization in the wild," in *Proc. ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., 2022, pp. 68–87.
[66] D. Kim, K. Wang, K. Saenko, M. Betke, and S. Sclaroff, "A unified framework for domain adaptive pose estimation," in *Proc. ECCV*, 2022, pp. 603–620.
[67] A. Juhong and C. Pintavirooj, "Face recognition based on facial landmark detection," in *Proc. 10th Biomed. Eng. Int. Conf. (BMEiCON)*, Aug. 2017, pp. 1–4.
[68] D. Anghelone, S. Lannes, V. Strizhkova, P. Faure, C. Chen, and A. Dantcheva, "TFLD: Thermal face and landmark detection for unconstrained cross-spectral face recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–9.
[69] L. Huang, Y. Liu, L. Chen, E. Z. Chen, X. Chen, and S. Sun, "Robust landmark-based stent tracking in X-ray fluoroscopy," in *Proc. ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., 2022, pp. 201–216.
[70] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," in *Proc. Conf. Workshop. Neur. Inf. Process. Syst. (NIPS)*, 2022, pp. 38571–38584.
[71] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2235–2245.
[72] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2129–2138.
[73] J. McCouat and I. Voiculescu, "Contour-hugging heatmaps for landmark detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20565–20573.
[74] S. Wachter, B. Mittelstadt, and C. Russell, "Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law," *West Virginia Law Rev.*, vol. 123, no. 3, pp. 1–51, Jan. 2021.

[75] N. Martinez, M. Bertran, and G. Sapiro, "Minimax Pareto fairness: A multi objective perspective," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6755–6764.
[76] J. Z. Liu, K. D. Dvijotham, J. Lee, Q. Yuan, B. Lakshminarayanan, and D. Ramachandran, "Pushing the accuracy-group robustness frontier with introspective self-play," in *Proc. ICLR*, 2023, pp. 1–10.
[77] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[78] V. V. Ramaswamy, S. S. Y. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9297–9306.
[79] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Duong, and M. Ghassem, "COVID-19 image data collection: Prospective predictions are the future," *Mach. Learn. Biomed. Imag.*, vol. 1, pp. 1–38, Dec. 2020.
[80] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1820–1828.
[81] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4352–4360.
[82] X. Lin, S. Kim, and J. Joo, "FairGRAPE: Fairness-aware gradient pruning method for face attribute classification," in *Proc. ECCV*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., 2022, pp. 414–432.
[83] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization," in *Proc. ICLR*, 2020, pp. 1–10.
[84] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., 2018, pp. 466–481.
[85] H. Nam, H. Lee, J. Park, W. Yoon, and D. Yoo, "Reducing domain gap by reducing style bias," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8686–8695.
[86] Y. Shi, J. Seely, P. H. Torr, N. Siddharth, A. Hannun, N. Usunier, and G. Synnaeve, "Gradient matching for domain generalization," in *Proc. ICLR*, 2022, pp. 1–10.
[87] S. Bucci, M. R. Loghmani, and T. Tommasi, "On the effectiveness of image rotation for open set domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 422–438.
[88] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu, "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *Ann. Statist.*, vol. 41, no. 5, pp. 2263–2291, Oct. 2013.
[89] Z. Wang, X. Dong, H. Xue, Z. Zhang, W. Chiu, T. Wei, and K. Ren, "Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10369–10378.
[90] W. Alghamdi, H. Hsu, H. Jeong, H. Wang, P. W. Michalak, S. Asoodeh, and F. Calmon, "Beyond adult and COMPAS: Fair multi-class prediction via information projection," in *Proc. NeurIPS*, 2022, pp. 38747–38760.
[91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[92] T. M. Dinh, A. T. Tran, R. Nguyen, and B.-S. Hua, "HyperInverter: Improving StyleGAN inversion via hypernetwork," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11379–11388.
[93] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

LEONARDO IURADA (Graduate Student Member, IEEE) received the B.S. degree in electronic and computer engineering from the University of Trieste and the M.S. degree in computer engineering from the Polytechnic of Turin, Italy, where he is currently pursuing the degree with the National Ph.D. Program in Artificial Intelligence. His work focuses on the development of trustworthy and efficient models for visual recognition.

SILVIA BUCCI received the Ph.D. degree from the Polytechnic of Turin, in 2023. Her research interests include open-set and generalizable learning models with applications in computer vision and robotics. She was awarded of the Blaceflor Foundation Scholarship, in 2022, and received the ICIAP Best Paper Award, in 2019. She served as a Reviewer for several top conferences and journals, such as CVPR, ECCV, ICCV, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *CVIU*, *IJCV*, and RAL; and was also a co-organizer of the Woman in Computer Vision Workshop at CVPR 2022.

TIMOTHY M. HOSPEDALES (Senior Member, IEEE) is currently a Professor with the School of Informatics, The University of Edinburgh. He is also a Principal Scientist and the Program Director of Machine Learning and Data Intelligence with the Samsung AI Centre, Cambridge. His research interests include machine learning and computer vision, with a focus on lifelong learning, broadly defined to include multi-domain/multi-task learning, domain adaptation, transfer learning, and meta-learning. In these areas, he has coauthored numerous papers in major venues, including CVPR, ICCV, ECCV, ICML, ICLR, NeurIPS, and AAAI. He is also an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He serves as the Area Chair for several major events, including ICCV, CVPR, ECCV, AAAI, and ACL.

TATIANA TOMMASI received the Ph.D. degree from EPFL Lausanne, in 2013. She is currently an Associate Professor with the Department of Control and Computer Engineering, Polytechnic of Turin, Italy, and the Director of the ELLIS Unit, Turin. She spent postdoctoral periods in Belgium and USA, before covering the role of an Assistant Professor with Sapienza University, Rome, Italy. She has published over 50 papers at top conferences and journals in machine learning and computer vision. She has a strong record in theoretically grounded algorithms for learning from images with robotics, medical, and human–machine interaction applications. She pioneered the area of transfer learning in computer vision and has extensive experience in domain adaptation, generalization, multimodal, and open-set learning. She is also an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement