E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2024)07-1814-20

论文引用格式:Wang M, Deng W H and Su S. 2024. Review on fairness in image recognition. Journal of Image and Graphics, 29(07):1814-1833(王玫, 邓伟洪, 苏森. 2024. 面向图像识别的公平性研究进展. 中国图象图形学报, 29(07):1814-1833)[DOI:10.11834/jig.230226]

面向图像识别的公平性研究进展

王玫[1],邓伟洪[2*],苏森[2]

1. 北京师范大学人工智能学院,北京　100875; 2. 北京邮电大学人工智能学院,北京　100876

摘　要:在过去的几十年里,图像识别技术经历了迅速发展,并深刻地改变着人类社会的进程。发展图像识别技术的目的是通过减少人力劳动和增加便利来造福人类。然而,最近的研究和应用表明,图像识别系统可能会表现出偏见甚至歧视行为,从而对个人和社会产生潜在的负面影响。因此,图像识别的公平性研究受到广泛关注,避免图像识别系统可能给人们带来的偏见与歧视,才能使人完全信任该项技术并与之和谐相处。本文对图像识别的公平性研究进行了全面的梳理回顾。首先,简要介绍了偏见3个方面的来源,即数据不平衡、属性间的虚假关联和群体差异性;其次,对于常用的数据集和评价指标进行汇总;然后,将现有的去偏见算法划分为重加权(重采样)、图像增强、特征增强、特征解耦、度量学习、模型自适应和后处理7类,并分别对各类方法进行介绍,阐述了各方法的优缺点;最后,对该领域的未来研究方向和机遇挑战进行了总结和展望。整体而言,学术界对图像识别公平性的研究已经取得了较大的进展,然而该领域仍处于发展初期,数据集和评价指标仍有待完善,针对未知偏见的公平性算法有待研究,准确率和公平性的权衡困境有待突破,针对细分任务的独特发展趋势开始呈现,视频数据的去偏见算法逐渐受到关注。

关键词:公平性;偏见;去偏见学习;图像识别;深度学习

Review on fairness in image recognition

Wang Mei[1], Deng Weihong[2*], Su Sen[2]

1. School of Artificial Intelligence, Beijing Normal University, Beijing　100875, China

2. School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing　100876, China

Abstract: In the past few decades, image recognition technology has undergone rapid developments and has been integrated into people's lives, profoundly changing the course of human society. However, recent studies and applications indicate that image recognition systems would show human-like discriminatory bias or make unfair decisions toward certain groups or populations, even reducing the quality of their performances in historically underserved populations. Consequently, the need to guarantee fairness for image recognition systems and prevent discriminatory decisions to allow people to fully trust and live in harmony has been increasing. This paper presents a comprehensive overview of the cutting-edge research progress toward fairness in image recognition. First, fairness is defined as achieving consistent performances across different groups regardless of peripheral attributes (e.g., color, background, gender, and race) and the reasons for the emergence of bias are illustrated from three aspects. 1) Data imbalance. In existing datasets, some groups are overrepresented and others are underrepresented. Deep models will facilitate optimization for the overrepresented groups to boost

收稿日期:2023-04-18;**修回日期**:2023-08-09;**预印本日期**:2023-08-16

*通信作者:邓伟洪　whdeng@bupt.edu.cn

基金项目:国家自然科学基金项目(62306043, 62236003);中国博士后科学基金资助项目(2022M720517)

Supported by:National Natural Science Foundation of China (62306043, 62236003); China Postdoctoral Science Foundation (2022M720517)

the accuracy, while the underrepresented ones are ignored during training. 2) Spurious correlations. Existing methods continuously capture unintended decision rules from spurious correlations between target variables and peripheral attributes, failing to generalize the images with no such correlations. 3) Group discrepancy. A large discrepancy exists between different groups. Performance on some subjects is sacrificed when deep models cannot trade off the specific requirements of various groups. Second, datasets (e. g. , Colored Mixed National Institute of Standards and Technology database (MNIST), Corrupted Canadian Institute for Advanced Research-10 database (CIFAR-10), CelebFaces attributes database (CelebA), biased action recognition (BAR), and racial faces in the wild (RFW)) and evaluation metrics (e. g. , equal opportunity and equal odds) used for fairness in image recognition are also introduced. These datasets enable researchers to study the bias of image recognition models in terms of color, background, image quality, gender, race, and age. Third, the debiased methods designed for image recognition are divided into seven categories. 1) Sample reweighting (or resampling). This method simultaneously assigns larger weights (increases the sampling frequency) to the minority groups and smaller weights (decreases the sampling frequency) to the majority ones to help the model focus on the minority groups and reduce the performance difference across groups. 2) Image augmentation. Generative adversarial networks (GANs) are introduced into debiased methods to translate the images of overrepresented groups to those of underrepresented groups. This method modifies the bias attributes of overrepresented samples while maintaining their target attributes. Therefore, additional samples are generated for underrepresented groups, and the problem of data imbalance is addressed. 3) Feature augmentation. Image augmentation suffers from model collapse in the training process of GANs; thus, some works augment samples on the feature level. This augmentation encourages the recognition model to produce consistent predictions for the samples before and after perturbing and editing the bias information of features, making it impossible for the model to predict target attributes based on bias information and thus improving model fairness. 4) Feature disentanglement. This method is one of the most commonly used for debiasing, which removes the spurious correlation between target and bias attributes in the feature space and learns target features that are independent of bias. 5) Metric learning. This method utilizes the power of metric learning (e. g. , contrastive learning) to encourage the model to make predictions based on target attributes rather than bias information to promote pulling the same target class with different bias class samples close and pushing the different target classes with similar bias class samples away in the feature space. 6) Model adaptation. Some works adaptively change the network depth or hyperparameters for different groups according to their specific requirements to address group discrepancy, which improves the performance on underrepresented groups. 7) Post-processing. This method assumes black-box access to a biased model and aims to modify the final predictions outputted by the model to mitigate bias. The advantages and limitations of these methods are also discussed. Competitive performances and experimental comparisons in widely used benchmarks are summarized. Finally, the following future directions in this field are reviewed and summarized. 1) In existing datasets, bias attributes are limited to color, background, image quality, race, age, and gender. Diverse datasets must be constructed to study highly complex biases in the real world. 2) Most of the recent studies dealing with bias mitigation require annotations of the bias source. However, annotations require expensive labor, and multiple biases may occasionally coexist. Mitigation of multiple unknown biases must still be fully explored. 3) A tradeoff dilemma exists between fairness and algorithm performance. Simultaneously reducing the effect of bias without hampering the overall model performance is challenging. 4) Causal intervention is introduced into object classification to mitigate bias, while individual fairness is proposed to encourage models to provide the same predictions to similar individuals in face recognition. 5) Fairness on video data has also recently attracted attention.

Key words: fairness; bias; debiased learning; image recognition; deep learning

0 引 言

图像识别是计算机视觉的基本问题之一，旨在利用计算机对图像进行处理、分析和理解，让计算机具有识别图像中出现的人物、物体、动作等的能力。早期的传统方法通常基于手工设计的特征进行识别(Sánchez等，2013；Chapelle等，1999)。然而，识别任务的准确性在很大程度上取决于特征的设计，不同类别样本之间的高相似性和同一类别样本的高可变

性使得传统方法无法取得较好的性能。近年来，得益于大规模数据、深度学习算法、基础硬件资源这三驾马车，基于深度学习的图像识别技术（Krizhevsky等，2017；He等，2016）通过多层非线性变换对高复杂性数据进行建模，使用深度网络提取的特征来替代手工设计的特征，取得了巨大的进步。相比于人类，机器对机械性工作的容忍度高（Danziger等，2011）且可利用大量数据进行高效的分析与判断。自此，基于深度学习的图像识别技术持续发展，产业化和商业化进程不断提速，并越来越多地应用于医疗诊断、行政执法等高风险决策中。

然而，在显著提升人类福祉的同时，人们发现现有图像识别系统可能会表现出偏见甚至歧视行为，从而对个人和社会产生潜在的负面影响。例如，尼康相机的眨眼警告功能往往会误判亚洲人一直在眨眼；亚马逊的人脸识别工具错误地将28名美国国会议员认成了罪犯，尤其黑人的错误率高达39%；在胸部X线影像智能诊断系统中，女性患者、20岁以下的患者和西班牙裔患者的诊断率偏低，算法易将患有疾病的个体标记为健康个体，从而延误病情（Seyyed-Kalantari等，2021）。算法的偏见可能导致某些群体受到不公正对待，带来严重的伦理道德问题；同时，偏见也会影响算法的泛化性，导致算法在特定群体或场景下性能下降，降低服务质量，甚至威胁人民的生命财产安全。

在此背景下，公平性的概念应运而生。公平性是指算法在决策过程中不存在因其固有或后天的属性所引起的偏见或偏爱（Saxena等，2019）。作为可信人工智能的重要组成部分，公平性已经成为全球共识，并受到国际各界的重视。2018年12月，欧洲联盟委员发布《可信人工智能的伦理指南草案》，其中的“正义原则”要求开发者和实施者需要确保个体和少数群体不受偏见和歧视。2019年6月，美国国家科学技术委员会更新《国家人工智能研究与发展战略规划》，强调通过设计提高公平性、透明度和问责性，建设符合道德伦理的人工智能。同月，我国新一代人工智能治理专业委员会发布《新一代人工智能治理原则——发展负责任的人工智能》，指出人工智能发展各方面应遵循公平公正原则，“在数据获取、算法设计、技术开发、产品研发和应用过程中消除偏见和歧视”。

为响应这一呼吁，学术界对于图像识别算法的公平性研究也逐渐兴起。美国计算机学会（Association for Computing Machinery，ACM）于2018年开始专门设立公平性、问责制和透明度会议（ACM Conference on Fairness，Accountability，and Transparency，FAccT），研讨交叉领域的公平性问题。与此同时，包括计算机视觉国际会议（International Conference on Computer Vision，ICCV）（https://htcv-iccv2021.github.io/）、国际先进人工智能会议（Association for the Advancement of Artificial Intelligence，AAAI）（https://aibsdworkshop.github.io/2022/index.html）和计算机视觉与模式识别国际会议（International Conference on Computer Vision and Pattern Recognition，CVPR）（https://fadetrcv.github.io/2022/）在内的多个人工智能重要国际会议也专门设置研究专题讨论公平图像识别。围绕该问题，研究者们纷纷构建多种多样的数据集来研究图像识别对颜色、纹理、图像质量和种族等的偏见，提出不同指标来量化图像识别模型的公平性程度，并设计新颖的去偏见算法来缓解不公平问题。

近年来，一些研究者发表了公平性的相关综述论文。但是已有论文一般都针对机器学习或其他人工智能领域而非聚焦于图像识别。例如，Mehrabi等人（2021）、邓蔚等人（2020）和刘文炎等人（2021）均针对机器学习的公平性算法进行了总结和分析，由于其覆盖范围较广，对图像识别公平性算法的总结比较宏观。Wang等人（2023）和Sun等人（2019）分别对推荐系统的公平性和自然语言处理的性别偏见问题进行了讨论。在识别领域，Singh等人（2022）仅关注于人脸分析的公平性，如人脸属性识别和人脸识别。

不同于先前的综述研究，本文力争从数据集、评价指标与去偏见算法等方面对图像识别的公平性研究进行系统论述，归纳总结已有的技术与进展，并讨论未来的研究挑战。

1 图像识别公平性概述

1.1 公平性和去偏见的定义

在图像识别模型中，将待识别的属性称为目标属性，并将某些其他因素，如颜色、背景、纹理、性别、种族和年龄等，称为偏见属性（或敏感属性）。图像识别的偏见是指图像识别模型会对具有不同偏见属性的群体进行区别对待，并在目标属性的识别中对

不同群体显示出不同的准确率。例如，图像分类模型会对背景产生偏见，它能够成功地识别出沙漠中的骆驼，却无法识别出绿洲中的骆驼；长发识别系统会对性别产生偏见，它能够成功地识别出女性的长发，却无法识别出男性的长发。因此，图像识别公平性旨在消除识别模型对颜色、背景、性别和种族等因素的偏见，确保模型能够在具有不同偏见属性的群体上都达到一致的性能。

公平性问题属于社会学和哲学的概念，当具体场景、社会需求和文化背景等多方面因素不同时，人们对于公平的理解和定义也随之不同。狭义上看，公平性和去偏见并没有严格的区别。在图像识别中，该领域尚处于发展初期，很多问题尚未形成定论。但从现有文献（Mehrabi 等，2021）可以看出，大部分工作对于这两者并没有进行区分，并认为公平性就是算法不存在因其固有或后天的属性所引起的偏见。

广义上看，去偏见主要只关注于结果公平（一致的性能），而公平性追求的不仅是结果公平，还包括起点公平和过程公平。然而，在图像识别领域，完全的公平暂时很难达到，例如在人脸识别中，即使使用相同的数据量和相同的算法来训练模型分别识别黑人和白人，黑人的识别率还是比白人低（Wang 等，2019a）。因此，在现阶段，图像识别领域的公平性主要还是关注于结果公平，即避免算法根据偏见信息做出错误决策，改善算法在弱势群体上的表现，并达到一致的性能。这是因为，在某些涉及道德伦理的问题上，如种族、性别和年龄等偏见，结果公平具有一定必要性，即使需要在一定程度上牺牲多数人的利益。例如，美国司法机构使用替代性制裁犯罪矫正管理剖析软件（correctional offender management profiling for alternative sanctions，COMPAS）来预测罪犯再犯罪的概率，然而该系统存在严重的种族偏见，两年内没有再犯罪的黑人被错误地归类为高风险的几率是白人的两倍。这种具有差异性、歧视性的结果势必会引起强烈的社会反响，保证一致的性能远比保证白人的高性能重要。此外，为了避免一味追求公平而过多牺牲性能的问题，现有很多工作（Iurada 等，2023；Zhang 和 Sang，2020）也开始提出需要兼顾公平性和准确率，综合两者指标来共同评价算法。

本文没有明确区分公平性和去偏见，因此，如果没有特别说明，在本文中两者可以等价。

1.2　偏见来源

现有文章（Nam 等，2020；Wang 等，2019a）表明，图像识别的偏见主要来源于数据不平衡、属性间的虚假关联和群体差异性 3 个方面，如图 1 所示。

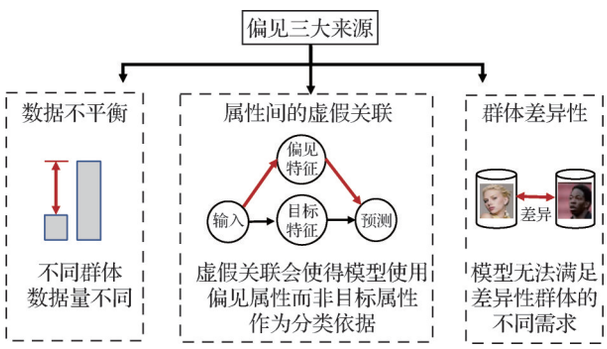


图 1　偏见的三大来源

Fig. 1　Three sources of bias

1）数据不平衡。现有大多数图像识别模型都是在大型、有标注的数据集上进行训练的。例如，图像分类模型通常在拥有超过 1 400 万幅标注图像的 ImageNet（Russakovsky 等，2015）上进行训练；人脸识别模型通常在拥有超过 1 000 幅标注图像的 MS-Celeb-1M（Guo 等，2016）上进行训练。研究人员通常使用特定的查询术语从谷歌图像或百度图像等网站上爬取图像，然后利用众包等方式对数据集进行标注。然而，这种方式可能会在无意中产生有偏见的数据（Zou 和 Schiebinger，2018）。在数据集中，某些群体的图像数量较多，而某些群体的图像数量相对较少。当不同群体的数据量不均衡时，以最小化整体误差为目标函数的模型会优先拟合多数群体以减小误差，而忽略少数群体，从而使得模型在不同群体上性能失衡。例如，Esteva 等人（2017）利用 129 450 幅图像训练深度模型以识别皮肤癌，然而只有不到 5% 的图像来自于深肤色人群，因此，该算法在不同肤色人群上存在性能差异。

2）属性间的虚假关联。可靠的识别算法应该依据正确的因果关系进行推理，例如数字识别模型应该依据数字的形状进行分类。然而，传统深度学习模型的训练大多以结果为导向，因此，算法往往专注于预测结果而不是理解因果关系。缺乏正确引导的深度学习模型会从多数群体样本中进行学习，在此过程中，目标属性和偏见属性间会产生虚假关联。因此，训练得到的模型会错误地依据偏见属性而非目标属性进行识别。同时，这种虚假的关联性往往

无法普适于所有样本，从而降低模型在少数群体上的性能，最终加剧偏见。例如，在数字识别任务Colored MNIST（Mixed National Institute of Standards and Technology database）的训练集中，大多数的“0”都是红色，大多数的“1”都是黄色。Nam等人（2020）发现，相比于形状，深度网络更容易学习到颜色信息。同时，Hong和Yang（2021）也通过实验发现，在特征空间中，不同的数字会根据颜色而非形状进行聚类。因此，他们均认为模型在训练的时候会捕捉到形状与颜色之间的相关性，这种虚假关联会使得模型使用更容易学习的颜色信息来代替形状作为分类依据进行分类。然而，实验证明，在测试时，该识别模型无法正确地对其他颜色的“0”或“1”进行识别。

3）群体差异性。不同群体的数据往往存在差异性，在某些特殊任务（如人脸识别）中，它们可能对于模型设计和模型训练具有不同的需求。当模型无法满足不同群体乃至不同样本的具体需求时，则会导致模型在某些群体上性能不佳，这也是产生偏见的原因之一。例如，Klare等人（2012）发现，在排除了训练集数据偏见的影响的情况下，使用不需训练的非深度学习算法得到的人脸识别模型仍在女性、深肤色和年轻群体上性能较差；Wang等人（2019a）证明了即使使用种族平衡的数据集训练的深度人脸识别模型在有色人种上的性能仍然不如白人，他们认为深肤色人群在黑暗环境下的图像对比度较低，因此，深肤色人群对识别模型有着更高的要求。

1.3 细分任务上的公平性

图像识别的细分任务有很多，如图像分类、人脸识别和行人重识别等。虽然这些任务都会因为某些偏见属性的存在，使得模型在训练过程中发生倾斜，最终在测试时，对某些群体样本表现出较差的性能。然而，不同任务的公平性问题存在着一些区别。

偏见的来源和机理略有不同。在图像分类中，除了数据不平衡外，偏见还由属性间虚假关联导致。每个目标类别都会与一个偏见类别高度相关，例如，在Colored MNIST中，不同数字都会对应不同的颜色（10类数字对应10种颜色），从而使得模型仅靠颜色就能对数字进行分类，使得模型产生偏见。而在人脸识别和行人再识别中，往往不存在目标类别和偏见类别一一对应的情况，模型不能仅靠偏见信息完成识别，例如，在人脸识别中，身份和种族并不是一一对应（上万个身份对应4个种族），因此不能仅靠种族对身份进行识别。在这些任务中，偏见主要来源于不同的群体（如种族、年龄、性别）彼此之间存在的分布差异性，使得适应性不佳的模型产生偏见，且不同群体的训练样本数量不同进一步放大了偏见，拟合了多数群体的模型在少数群体上性能不佳。

目标属性和偏见属性的耦合程度不同。相对于图像分类和行人重识别来说，人脸识别中的偏见属性和目标属性耦合程度较强。例如，在图像分类数据集Colored MNIST中，虽然颜色（偏见）和形状（目标）之间被构造出关联性，但在现实生活中，这两者之间并没有固有联系。然而，在人脸识别中，性别/种族和身份之间的关联性是固有的、天生的，甚至性别和种族这些偏见是人脸面部信息的一部分。因此，对于图像分类和行人重识别来说，特征解耦是一个较为可行的方法，但是对于人脸识别来说，将性别/种族与身份分离开来较为困难，且会造成目标任务识别率的下降（Gong等，2020）。

2 图像识别公平性数据集

在以数据驱动的深度学习时代，数据集是算法研究的基础。近年来，业界逐渐发布了一些图像识别的公平性数据集，以研究图像识别算法对于颜色、纹理、背景和群体（性别/年龄/种族）的偏见。表1从规模、目标属性和偏见属性等方面对常用的图像识别公平性数据集进行了归纳。

Colored MNIST数据集（Kim等，2019a；Nam等，2020）可用于研究数字识别模型对颜色的偏见。该数据集由对MNIST数据集（LeCun等，1998）中的灰度数字进行重新着色改造而成。首先，将10种不同的颜色分别分配给10个数字类别作为它们的平均颜色，以建立数字和颜色两个属性间的相关性；然后，对于每个训练图像，从相应平均颜色的正态分布中随机采样一种颜色，对数字进行着色。通过控制该正态分布的方差可以调整数据中的颜色偏见程度，方差越小，偏见程度越大。例如，Kim等人（2019a）将方差设置为{0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05}；Nam等人（2020）将方差设置为{0.05, 0.02, 0.01, 0.005}。对于每个测试图像，则从10种预定义的颜色中随机选择一种平均颜色来构造无偏测试集，并遵循与训练图像相同的着色协议。数据集链接为：https://github.com/alinlab/LfF。

Multi-Color MNIST数据集(Li等,2022)可用于研究数字识别模型对多种颜色的偏见。Li等人(2022)认为现实世界中的识别模型可能会受到多种偏见的影响,因此,他们为每个图像都添加左右两种不同的背景颜色,使数字与两种背景颜色同时产生相关性。链接为:https://github.com/zhihengli-UR/DebiAN。

Biased MNIST数据集(Shrestha等,2022a)对MNIST数据集进一步添加多重偏见属性,如图2(a)所示。具体而言,数据集中的图像由5×5的网格组成,数字放置在其中一个网格内,并与6种偏见属性产生相关性,包括:1)数字大小/比例(数字占据的单元格数量);2)数字颜色;3)背景纹理类型;4)背景纹理颜色;5)共存的字母;6)共存字母的颜色。其他设置均与Multi-Color MNIST数据集类似。该数据集链接为:https://github.com/erobic/bias-mitigators。

Corrupted CIFAR-10(Canadian Institute for Advanced Research-10 database)数据集(Nam等,2020)是通过对CIFAR-10数据集(Krizhevsky,2009)进行改造而来,用于研究图像分类模型对图像质量的偏见。遵循Hendrycks和Dietterich(2019)所提出的协议,Nam等人(2020)通过给10种不同类别的图像添加10种相应的损坏,建立图像类别与图像质量的相关性,从而构建两个公平性数据集Corrupted CIFAR-10[1]和Corrupted CIFAR-10[2]。其中,损坏类型分别为{对比度、亮度、饱和度、弹性形变、JPEG压缩、像素化、雪、雾、霜、飞溅}和{高斯噪声、散粒噪声、脉冲噪声、斑点噪声、高斯模糊、散焦模糊、玻璃模糊、运动模糊、缩放模糊、原图}。数据集链接为:https://github.com/alinlab/LfF/tree/master/data。

9-Class ImageNet数据集(Bahng等,2020)可用于研究图像分类模型对纹理的偏见,如图2(c)所示。该数据集是ImageNet(Russakovsky等,2015)的一个子集,包含狗、猫、青蛙、乌龟、鸟、灵长类动物、鱼、螃蟹和昆虫9个超类(Ilyas等,2019),每个超类的样本均与背景纹理具有相关性,例如模型会通过背景中水的纹理来识别乌龟。Bahng等人(2020)平衡了每个超类中各子类图像的比例,以消除其他因素的影响,从而关注纹理偏见对模型预测的作用。该数据集生成方法的链接为:https://github.com/clovaai/rebias/blob/master/datasets/imagenet.py。

BAR(biased action recognition)数据集(Nam等,2020)是真实世界的动作识别数据集,共有6个动作类别,包含1 941个训练样本和654个测试样本,其中背景位置为偏见属性,如图2(b)所示。Nam等人(2020)分别从imSitu(Yatskar等,2016)、Stanford 40 Actions(Yao等,2011)和谷歌图像网站中收集图像,并挑选6类动作—背景位置相关的图像构建训练集,使得动作识别模型仅靠背景位置即可对图像进行分类。动作—背景位置对应关系分别为:攀岩—岩壁、潜水—水下、钓鱼—水面、赛车—赛道、投掷—操场和撑杆跳—天空。测试集则包含动作—背景位置无关的无偏图像,如在雪山上攀岩的图像。数据集链接为:https://github.com/alinlab/BAR。

bFFHQ(flickr-faces-high-quality database)数据集(Kim等,2021)基于Flickr Faces HQ数据集(Karras等,2019)改造而来。该数据集包含分辨率为1 024×1 024像素的人脸图像,且在年龄(目标属性)和性别(偏见属性)之间具有很强的相关性。训练集由19 200幅图像组成,其中,“年轻”(即10～29岁)与“女性”高度相关,“老年”(即40～59岁)则与“男性”高度相关。测试集包含年龄—性别无关的2 000幅图像以评估公平性。数据集链接为:https://github.com/zhihengli-UR/DebiAN/blob/main/datasets/bffhq.py。

CelebA(CelebFaces attributes database)数据集(Liu等,2015)是人脸属性识别的常用数据集,每个样本都标有40个属性,如图2(d)所示。该数据集由202 599幅人脸图像组成,其中,162 770幅图像用于训练,19 867幅图像用于测试。2020年,Nam等人(2020)对CelebA提出新的实验设置,以便用于公平性算法的研究。他们将发色和浓妆作为待识别的目标属性,并将性别作为偏见属性。由于大多数拥有金发并带有浓妆的图像都是女性,因此模型在发色和浓妆的预测上,会对不同性别群体产生偏见。随后,在其他的公平性研究中(Park等,2022;Seo等,2022;Zhang等,2022),CelebA的分类目标又扩展到了吸引力、微笑、大鼻子等属性上。该数据集的链接为:https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html。

UTKFace数据集(Zhang等,2017)是一个拥有20 K幅人脸图像的数据集,包含年龄(0～116岁)、性别(男和女)和种族(白人、黑人、亚洲人、印度人和其他)标签。Jung等人(2022)将其引入公平性研究中,并将种族和年龄分别作为偏见属性和目标属性。其中,年龄划分为3类:0～19岁、20～40岁和40岁以上;种族划分为4类:白人、黑人、亚洲人和印度人。

与此类似，Park等人（2021）、Hong和Yang（2021）则将性别作为偏见属性，年龄和种族作为目标属性。其中，年龄划分为两类：年轻和老年；种族划分为两类：白人和其他。数据集链接为：https://susanqq.github.io/UTKFace/。

IMDB（internet movie database）数据集（Rothe等，2018）是一个人脸图像数据集，包含来自20 284位名人的460 723幅图像，并对年龄和性别进行了标注。Kim等人（2019a）清洗并过滤掉标注错误的图像，将剩余的112 340幅图像引入公平性的研究中，其中，性别作为目标属性，年龄作为偏见属性。该数据集包含3个子集：EB1（0～29岁女性和40岁以上男性）、EB2（40岁以上女性和0～29岁男性）和测试集（性别—年龄无关的无偏图像）。模型在EB1上训练并在EB2和测试集上测试；反之亦然。链接为https://github.com/feidfoe/learning-not-to-learn/tree/master/dataset/IMDB。

Fairface数据集（Kärkkäinen和Joo，2021）是一个种族、年龄和性别均平衡分布的人脸图像数据集，包含108 501幅人脸图像，并对种族、年龄和性别进行了标注。该数据集的提出是为了消除种族、年龄和性别识别中的数据偏见。数据集链接为：https://github.com/joojs/fairface。

RFW（racial faces in the wild）数据集（Wang等，2019a）是一个种族平衡的测试集，用于评估人脸识别模型对种族的偏见程度。该数据集包含白人、印度人、东亚人和黑人4个子集，每个子集包含约3 000个人的1万幅图像用于人脸验证，采用ROC曲线（receiver operator characteristic curve）和LFW（labeled faces in the wild）协议计算每个种族的人脸识别性能，利用不同种族的性能差异衡量偏见程度。链接为：http://www.whdeng.cn/RFW/index.html。

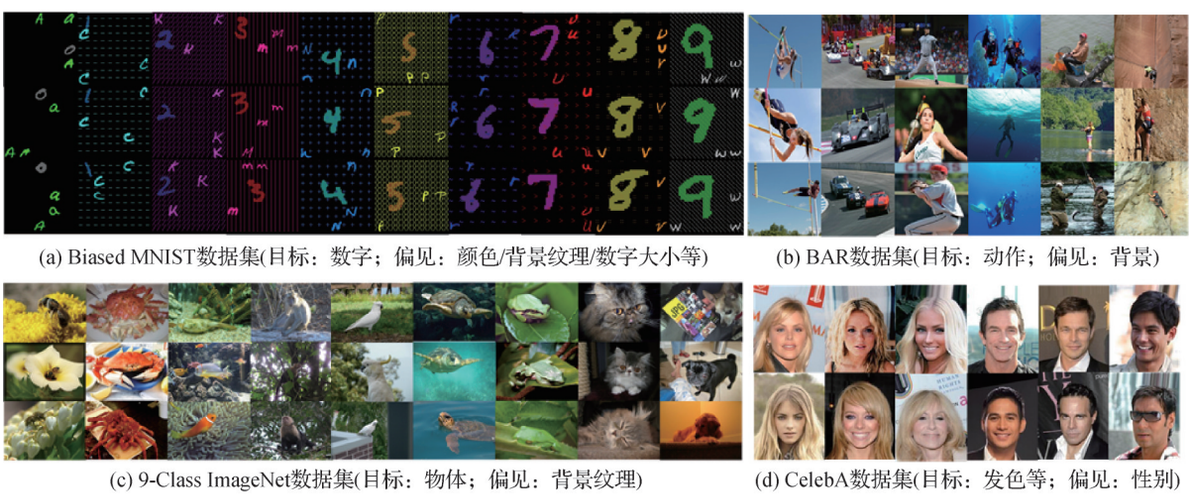
表1　常用的图像识别公平性数据集

Table 1　Commonly-used datasets for fair image recognition

数据集	发布机构	图像规模	目标属性	目标类别数	偏见属性
Colored MNIST（Kim等，2019a）	韩国科学技术院	60 K	数字	10	字体颜色
Multi-Color MNIST（Li等，2022）	美国罗切斯特大学	60 K	数字	10	背景颜色
Biased MNIST（Shrestha等，2022a）	美国罗切斯特理工学院	60 K	数字	10	颜色/纹理/背景/位置
Corrupted CIFAR-10（Nam等，2020）	美国加州大学伯克利分校	60 K	物体	10	图像质量
9-Class ImageNet（Bahng等，2020）	韩国大学	56.7 K	物体	9	背景纹理
BAR*（Nam等，2020）	美国加州大学伯克利分校	2 595	动作	6	位置
bFFHQ（Kim等，2021）	韩国科学技术院	21.2 K	年龄	2	性别
CelebA（Liu等，2015）	香港中文大学	202 599	浓妆/长发/吸引力	2	性别/年龄
UTKFace（Zhang等，2017）	美国田纳西大学	20 K	种族/性别/年龄	2	性别/种族
IMDB（Rothe等，2018）	瑞士苏黎世联邦理工学院	460 723	性别	2	年龄
Fairface*（Kärkkäinen和Joo，2021）	美国加州大学洛杉矶分校	108 K	性别/种族/年龄	2/7/9	年龄/性别/种族
RFW*（Wang等，2019a）	北京邮电大学	40 607	人脸身份	12K	种族
BUPT-Balancedface*（Wang和Deng，2020）	北京邮电大学	1.3 M	人脸身份	28K	种族
BUPT-Globalface*（Wang和Deng，2020）	北京邮电大学	2 M	人脸身份	38K	种族

注：带*表示针对公平性研究提出的数据集，不带*表示通过对传统图像识别数据集添加偏见改造而成。各数据集下载链接详见文内介绍。

BUPT-Balancedface 和 BUPT-Globalface 数据集(Wang 和 Deng，2020)是人脸识别领域的两个公平性训练集，包含白人、印度人、东亚人和黑人4个种族。其中，BUPT-Balancedface 包含来自 2.8 万个人的 130 万幅图像，每个种族大约都包含 7 000 个名人以消除现有人脸识别训练集中的种族偏见，平等地表征不同种族。BUPT-Globalface 包含 3.8 万人的 200 万幅图像，其种族分布与世界人口的真实分布大致相同，以模仿真实世界的偏见程度。数据集链接为：http://www.whdeng.cn/RFW/index.html。



(a) Biased MNIST数据集(目标：数字；偏见：颜色/背景纹理/数字大小等)　(b) BAR数据集(目标：动作；偏见：背景)

(c) 9-Class ImageNet数据集(目标：物体；偏见：背景纹理)　(d) CelebA数据集(目标：发色等；偏见：性别)

图 2　图像识别公平性数据集示例

Fig. 2　Examples of some fairness datasets in image recognition ((a) Biased MNIST dataset (target: digit, bias: color/ background texture/digit size); (b) BAR dataset (target: action, bias: background); (c) 9-Class ImageNet dataset (target: object, bias: background texture); (d) CelebA dataset (target: hair color, bias: gender))

3　图像识别公平性评价指标

为衡量公平性，机器学习领域提出了多样的公平性指标，包括几率均等(equalized odds，EOD)、机会均等(equal opportunity，EOP)、人口平价(demographic parity，DP)、条件统计奇偶公平性(conditional statistical parity，CSP)、感知公平性(fairness through awareness，FTA)、不感知公平性(fairness through unawareness，FTU)和反事实公平性(counterfactual fairness，CF)等。然而，图像识别中较为常用的公平性指标主要是识别率、几率均等、机会均等和识别率方差这4种。因此，本节只对这4种指标进行简要介绍，其他机器学习相关的公平性指标可参考 Mehrabi 等人(2021)的工作。

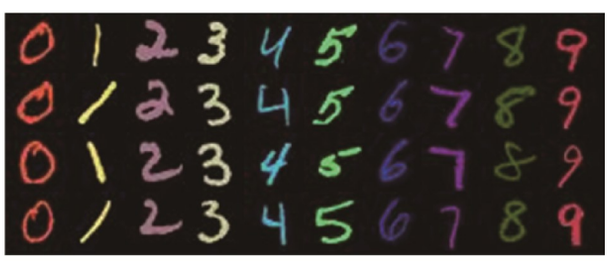
1)识别率。模型在无偏测试集上的识别率是较为常用的公平性指标(Nam 等，2020)。无偏测试集是指由目标与偏见无关的样本组成的集合，例如，在 Colored MNIST 中，每个数字的颜色都是完全随机的。该指标简单直观，能够展示模型的真实性能。此外，模型在偏见冲突样本上的识别率也可以用来衡量偏见。Nam 等人(2020)将偏见冲突样本定义为无法利用偏见属性对目标类别进行正确预测的样本；并将偏见对齐样本定义为可以利用偏见属性进行正确预测的样本。如图 3 所示，在 Colored MNIST 中，偏见对齐样本就是第 2 节中提到的利用平均颜色分布中采样的颜色进行着色的样本，而偏见冲突样本则利用其他(9 种)颜色分布中采样的颜色进行着色。当使用大量偏见对齐样本和少量偏见冲突样本进行训练时(两者的比例可以反映训练集的偏见程度)，模型会倾向于依靠偏见属性进行目标分类，因此，使用偏见冲突样本进行识别率计算可以展示模型的真实目标分类性能。

2)几率均等(EOD)。Hardt 等人(2016)利用几率均等来衡量公平性。该项指标指出，若模型的预测 $\hat{Y}$ 针对偏见属性 A 和目标属性 Y 满足几率均等，则在 Y 的条件下，A 和 $\hat{Y}$ 应该是独立的，即 $P\left(\hat{Y}=1|A=0,Y=y\right)=P\left(\hat{Y}=1|A=1,Y=y\right),y\in\{0,1\}$。这意味着，对于具有不同偏见属性的群体来说(如男性和女性)，正样本被正确分类且负样本被错误分类的概

率应该是相同的。换言之，几率均等定义指出，真阳率（true positive rate，TPR）和假阳率（false positive rate，FPR）在具有不同偏见属性的群体之间应该保持相等。因此，利用以下计算来衡量公平性，具体为

$$EOD = \frac{1}{2}\left[\left|TPR_{A=0} - TPR_{A=1}\right| + \left|FPR_{A=0} - FPR_{A=1}\right|\right] \tag{1}$$

式中，$TPR_{A=0}$表示在偏见属性为0的样本上的TPR，$FPR_{A=0}$表示在偏见属性为0的样本上的FPR，$TPR_{A=1}$表示在偏见属性为1的样本上的TPR，$FPR_{A=1}$表示在偏见属性为1的样本上的FPR。



(a) 偏见对齐样本

(b) 偏见冲突样本

图3　Colored MNIST中偏见对齐样本和偏见冲突样本示例

Fig. 3　Illustration of bias-aligned samples and bias-conflicting samples in Colored MNIST ((a) biased-aligned samples; (b) biased-conflicting samples)

3）机会均等（EOP）。Hardt等人（2016）提出的机会均等指出，若模型的预测$\hat{Y}$针对偏见属性A和目标属性Y满足机会均等，则应该满足在$Y=1$的条件下，A和$\hat{Y}$是彼此独立的，即$P\left(\hat{Y}=1|A=0,Y=1\right)=P\left(\hat{Y}=1|A=1,Y=1\right)$。这意味着，对于具有不同偏见属性的群体来说（如男性和女性），正样本被正确分类的概率应该是相同的。换言之，机会均等定义指出，真阳率在具有不同偏见属性的群体之间应该保持相等。因此，利用以下计算来衡量公平性，具体为

$$EOP = \left|TPR_{A=0} - TPR_{A=1}\right| \tag{2}$$

式中，$TPR_{A=0}$和$TPR_{A=1}$分别表示在偏见属性为0和1的样本上的TPR。

4）识别率方差。为了更好地描述具有不同偏见属性的群体之间的性能差异（超过两种偏见属性，如多种族、多年龄），Wang和Deng（2020）提出使用不同群体识别率的方差来衡量偏见程度。

4　图像识别公平性算法

随着计算机硬件的升级、相关训练数据集的扩充，深度学习飞速发展，并在图像识别领域取得巨大突破。然而，人们发现在视觉识别任务上已经超过人类的深度学习算法却存在着偏见与歧视。因此，针对深度学习的去偏见算法也逐渐受到人们的关注。在图像识别领域，现有的去偏见算法主要分为：重加权（重采样）、图像增强、特征增强，特征解耦、度量学习、模型自适应和后处理。

4.1　重加权（重采样）

样本重加权（重采样）是最早被提出，也是最常用的去偏见算法之一。在训练时，该算法通过为少数群体赋予更高的权重（增加少数群体的采样频率），并为多数群体赋予较低权重（降低多数群体的采样频率），提升模型对少数群体的关注程度，降低对多数群体的拟合，从而减小模型偏见。近年来，相关研究主要关注于如何生成合适的权重（确定合适的采样频率）以提升公平性。Nam等人（2020）认为只有当偏见属性比目标属性更容易被学习时，模型才会依赖虚假关联性进行预测，而这种依赖在训练的早期阶段最为突出。因此，在模型训练过程中，偏见对齐样本的损失函数值更小，而偏见冲突样本的损失函数值更大。他们利用样本的损失函数值对其进行重加权，从而提升对偏见冲突样本的关注程度。由于样本数量不平衡是导致模型偏见的原因之一，Park等人（2022）利用不同群体的样本数量来生成权重。在此基础上，Seo等人（2022）则提出可以同时考虑图像数量和损失函数值的大小。Kim等人（2022）综合考虑多个分类器的结果来检测偏见冲突样本，并利用样本预测难度来生成权重。Ahn等人（2023）和Zhao等人（2021）受到分布外（out-of-distribution，OOD）检测（Huang等，2021）的启发，观察到与分布外样本类似，偏见冲突样本的梯度模值较大，因此利用梯度模值对样本进行重加权。Li和Vasconcelos（2019）将样本权重作为一个可学习的变量，将偏见最小化等同于重加权后的数据在分类器上的损失与

标签不确定度之间的比例，并使用随机梯度下降（stochastic gradient descent，SGD）交替更新分类器参数和样本权重来减小偏见。Amini 等人（2019）将识别任务与变分自动编码器（variational autoencoders，VAE）融合以学习数据集中的潜在分布，然后在训练时自适应地使用学习的潜在分布来生成权重。Bruveris 等人（2020）依据不同群体的训练性能对样本进行重采样，从而减小人脸识别的偏见。

讨论：重加权或重采样的方法简单易行，且不会增加过多网络负担。然而，重加权的方法并没有在本质上解决问题，仅靠减小对少数群体的拟合，增加对多数群体的关注程度，达到的效果非常有限。另外，重采样的方法降低对多数群体图像的采样频率，甚至对其进行丢弃，大大降低了对图像的利用率，造成数据资源浪费，极可能在提升公平性的同时降低准确率。

4.2 图像增强

为缓解数据不平衡所带来的偏见问题，并增强样本的多样性，部分工作引入生成对抗网络（generative adversarial network，GAN）进行样本生成，以增加少数群体的图像数量。Kim 等人（2021）利用编码器分别生成目标特征和偏见特征，并交换任意两个偏见对齐样本和偏见冲突样本的偏见特征，利用生成对抗网络增广少数样本。同时，为了改善生成质量，他们利用类激活图（class activation map，CAM）来指导判别器重点关注对偏见影响较大的区域，从而在交换偏见的过程中生成更加真实和准确的图像。Ramaswamy 等人（2021）训练生成对抗网络生成真实的图像，并提出一种在生成对抗网络的潜在空间中扰动潜在向量的方法来更加高效、便捷地更改图像的偏见属性，从而平衡训练数据。Georgopoulos 等人（2021）基于自适应实例归一化（adaptive instance normalization，AdaIN）从网络的不同层次同时提取图像的多种属性，将提取的属性特征融合并注入解码器中，从而实现将不同图像的多种偏见属性转换迁移至其他图像中，最终生成偏见属性分布平衡的训练集以消除偏见。Ge 等人（2020）利用 starGAN（Choi 等，2018），在保持身份不变的情况下，改变图像的种族属性，从而扩增黑人图像以解决人脸识别中的种族偏见问题；Yucer 等人（2020）基于 CycleGAN（Zhu 等，2017）实现种族属性的转变。考虑到对抗样本可以通过微小的扰动来误导分类器的预测，例如将男性错误地分类为女性，因此，Zhang 和 Sang（2020）提出利用对抗样本来改变图像的偏见属性，增加少数群体的样本数量。他们以在线方式将目标属性分类器、偏见属性分类器和对抗样本生成器的训练过程耦合起来，从而确保对抗样本的泛化性和跨任务可迁移性，提升图像生成效果和去偏见效果。

讨论：基于生成对抗网络的图像增强可以有效地解决数据不平衡问题，通过从数据层面消除偏见避免了公平性和准确性的权衡困境问题。但是由于依赖生成器和判别器的对抗学习，生成对抗网络的训练极不稳定，并且容易出现模型坍塌等问题。

4.3 特征增强

鉴于图像增强的局限性，研究者们提出在特征层面进行增强的方法，通过对特征的偏见信息进行扰动和编辑，并约束分类器能正确分类增强后的样本，从而减小识别模型对于偏见属性的依赖性，提升公平性。Du 等人（2021）在不改变特征提取器的情况下，利用增强后的特征训练分类器，使其能够利用有偏见的特征得到公平的预测。具体来说，他们提取两个具有相同目标类别而不同偏见类别的样本的特征，并对两者特征进行加权平均来得到增强后的特征，通过约束增强后特征的预测分数与增强前预测分数的加权平均保持一致，从而使得分类器对于偏见特征不敏感。Chuang 和 Mroueh（2021）提出在特征流形上对具有相同目标类别而不同偏见类别的两个样本进行混合操作（MixUp）以实现特征增强，在以不同比例进行特征混合的情况下，利用平滑约束鼓励模型保持一致的预测结果，减小偏见对预测的影响。Nuriel 等人（2021）认为某些偏见（如纹理）可以归为风格信息，因此，基于自适应实例归一化（Huang 和 Belongie，2017），对同一批次内的两个样本特征进行风格信息（即底层的纹理信息）迁移，以改变样本的偏见属性而实现特征增强，并约束模型能够正确分类增强后的特征来消除图像分类模型对纹理的偏见。

讨论：相比于图像增强，特征增强的训练方法更加稳定。通过增加特征的多样性，能够在一定程度上解除偏见和目标信息的相关性。然而，增强后的特征语义性和解释性不强；且当前特征增强的方法仍然较简单，增强后的特征多样性相对不足。

4.4 特征解耦

特征解耦是最常用的去偏见算法，它在特征空间中解除目标属性和偏见属性的关联性，并学习到偏见无关的目标特征，从而使得模型的识别能够不受偏见的影响。Ragonesi等人(2021)引入互信息的思想，估计并减小目标特征和偏见标签之间的互信息，从而去除目标属性特征中的偏见信息。Zhu等人(2021)利用同时结合内容和局部结构的表示学习来增强互信息估计器，并减小目标特征和偏见特征之间的互信息。Sarhan等人(2020)分别对目标特征和偏见特征进行编码，并利用最大化熵来消除目标特征中的偏见信息，同时借助于变分推断(variational inference, VI)来约束两类特征的正交性，进一步进行特征解耦。Park等人(2021)认为，由于目标属性和偏见属性的高度耦合，无法将两者完全解耦。因此，他们将图像映射到目标、偏见和交互信息3个特征空间，并利用再编码机制提取交互特征中的有用信息，联合目标特征共同进行识别。

同时，一些研究将对抗学习(Tzeng等，2017；Ganin和Lempitsky，2015)引入公平性算法中，利用其将偏见特征从目标特征中去除，消除两者间的关联性。Alvi等人(2018)和Kim等人(2019a)优化一个辅助分类器来识别偏见属性，同时优化特征提取器来混淆辅助分类器，使其无法对偏见属性进行正确分类。通过辅助分类器和特征提取器的对抗学习来得到偏见无关的目标属性。Gong等人(2020)将人脸特征解耦为身份、年龄、性别和种族4种特征，使得人脸识别模型不受年龄、性别和种族偏见的影响。为鼓励特征间的独立性，他们利用对抗学习进行去相关，并约束4种特征的联合分布与各自边缘分布的乘积相等。Dhar等人(2021)利用对抗学习对预训练模型中提取到的人脸特征进行去偏，将其映射到偏见无关的特征空间。

讨论：特征解耦的方法能够解除偏见和目标间的相关性，生成偏见无关的、独立的目标特征，从而在本质上解决偏见问题。然而，特征解耦算法往往面临准确性和公平性的权衡困境。精确的特征解耦往往无法保证，若解耦程度较弱，则目标特征中仍然保留偏见信息，去偏见效果不佳；若解耦程度较强，则某些目标信息也会被错误地去除，从而影响目标分类的性能。同时，在某些特殊任务上，偏见属性也是目标属性的一部分。例如在人脸识别中，种族也是面部身份信息的一部分，如果直接暴力解耦两者特征，则必定会影响算法的准确度。

4.5 度量学习

为了使得模型根据目标属性而非偏见属性进行分类，一种最直观的想法就是利用度量学习使得具有相同目标属性和不同偏见属性的样本彼此靠近，具有不同目标属性和相同偏见属性的样本彼此远离。Hong和Yang(2021)发现在有偏见的模型的特征空间中，样本会根据偏见属性而非目标属性进行聚类。因此，为了避免产生这种错误的聚类，使得模型能够依据目标属性进行分类，他们基于SupCon(supervised contrastive loss)(Khosla等，2020)进行对比学习，并更改正样本对的设置，将相同目标类别而不同偏见类别的样本作为正对，以减小它们之间的距离。Park等人(2022)则更改负样本对的设置，将不同目标类别而相同偏见类别的样本作为负对，以增加它们之间的距离。Zhang等人(2023)首先利用AttGAN(attribute GAN)(He等，2019)对图像进行增强，在保持其他属性不变的情况下，使得生成的图像和原图具有不同的偏见属性；然后利用对比学习进行模型优化，将原图和生成的图像作为正对，将具有相同偏见属性的不同图像作为负对。Jung等人(2021)提出利用一个已有的、有偏见的模型作为教师，蒸馏出一个公平的学生模型。为了保证公平性，在蒸馏的同时，他们基于最大均值差异(maximum mean discrepancy, MMD)拉近学生模型的组条件特征分布和教师模型的组平均特征分布间的距离，使得学生模型中具有不同偏见属性的样本都向教师模型中同目标类别的特征靠近。

讨论：度量学习的思想较为直观，直接对特征空间进行约束，使得具有相同目标类别而不同偏见类别的样本距离更近，从而保证模型的目标分类与偏见属性无关。然而，该方法对于偏见标签的依赖性非常强，当样本的偏见标签缺失时，该方法则失效。

4.6 模型自适应

面对差异化的样本和群体时，单一的模型可能无法满足每个样本或群体的独特需求，从而使得模型在某些样本上性能较差，最终导致不公平性。因此，一些研究提出，提高模型对不同样本的适应能力能够在一定程度上满足每个样本和群体的去偏见需求。Shrestha等人(2022b)基于奥卡姆剃刀原理，提出使用统一深度(复杂度)的网络对所有样本进行学

习不利于模型的公平性。对于某些样本来说，有时较为简单的结构反而有利于网络忽略其虚假关联因素。因此，他们利用门机制自适应地为每个样本改变网络的深度。在人脸识别的种族偏见领域，Wang和Deng（2020）发现即使在样本数量平衡的情况下，训练得到的人脸识别模型中不同种族特征的类间可分性仍然不同。因此，他们提出对模型的超参数进行自适应调整以满足不同群体的需求。具体来说，提出了一种基于强化学习的种族平衡网络（reinforcement learning based race balance network，RL-RBN）（Wang和Deng，2020），引入强化学习对大间距损失函数（Deng等，2019；Wang等，2018）中的间距参数进行自适应搜索。该网络将间距参数的调整作为动作，将白人与其他种族类间可分性的差异作为状态，将群体间特征散度的平衡性作为奖励，从而控制模型的学习过程，使各群体具有相似的类间可分性。此外，他们还提出一种基于元学习的元平衡网络（meta balanced network，MBN）（Wang等，2022），将间距参数作为一个可学习的变量，利用内层和外层循环迭代地更新模型参数和间距参数。具体来说，在外层循环中，使用小型且平衡的元数据集来评估当前模型的偏见，并基于偏见感知损失在元数据上的梯度来自适应地为每个群体生成最优的间距参数。Gong等人（2021）为每个种族生成合适的卷积核掩码和通道注意力图，从而能更加灵活地激活不同的面部区域以用于识别，满足不同群体的训练需求。

讨论：模型自适应方法基于网络结构或超参数进行公平性学习，可以与其他图像层和特征层的去偏见方法相结合。然而，目前模型自适应仅在某些特殊任务上进行了初步探索，如人脸识别，它的普适性尚未得到验证。

4.7　后处理

基于后处理的去偏见算法（Hardt等，2016）是在不改变训练过程的前提下，对训练后的模型或模型的预测结果进行处理，以消除训练过程中残余的不公平性。Terhörst等人（2020a）提出用一个公平性驱动的神经网络分类器来代替相似性函数，用于比较两个人脸的特征。在该分类器的训练中，将公平性标准引入决策过程，迫使不同种族的分数分布相似，从而减少不同种族的性能差异。Terhörst等人（2020b）提出一种分数标准化方法，对训练样本进行聚类划分，根据聚类集合上的等错误率（equal error rate，EER）的阈值来动态调整人脸验证时的相似度分数，实现分数标准化以提升公平性。Wang等人（2020）提出，在推断时通过集成在具有不同偏见属性的群体上独立训练的识别模型，并应用先验偏移推断来消除偏见信息。Kim等人（2019b）使用一个小规模数据集来审计预训练模型的偏见程度，并对预训练模型进行后处理，利用乘性权重改善分类器，使得该分类器在具有不同偏见属性的群体上达到平衡的结果。

讨论：基于后处理的去偏见算法可以在不改变训练过程的情况下，实现模型的去偏见。对于黑盒模型来说，后处理技术是不得不采用的技术路径。同时，不需要重新训练的策略也节省了计算资源，提升了效率。然而，后处理是一种事后补救措施。虽然可以结合其他去偏见算法一起使用，以消除模型中残留的偏见和不公平性，但是某些训练数据或模型算法的固有偏见在本质上难以事后纠正或消除。因此，相比于其他方法，后处理的研究工作还比较局限。

4.8　主流数据集上的测试结果

图像识别的公平性研究仍处于发展初期，公平性的评估标准尚未统一。不同算法选择不同的数据集进行性能评估；即使在同一数据集上，所采用的评价指标也不尽相同；即使采用的数据集和评价指标相同，数据集的构造方式（如训练集中偏见对齐和偏见冲突样本的比例）也不一样。因此，现阶段很难对各类公平性算法的性能进行统一的对比和比较，本小节中仅列举了部分公平性算法在几个主流数据集上的实验结果，如表2—表5所示，以便大致了解当前算法的水平。其中，表2中的结果引用自Zhang等人（2022）和Park等人（2022）的工作；表3中的结果引用自Hwang等人（2022）的工作；表4和表5中的结果引用自Wang等人（2022）和Gong等人（2021）的工作。

表2展示了6种去偏见方法在Colored MNIST、Corrupted CIFAR-10和bFFHQ的无偏测试集上的识别率，以衡量不同方法对颜色、图像质量和性别偏见的去偏见效果。从第2列可以看出，当不使用任何去偏见方法时，模型的无偏识别率随着数据集偏见程度的增大（偏见冲突样本比例的减小）而降低。在这6种去偏见算法中，LfF（learning from failure）（Nam

等，2020）采用重加权来得到无偏模型；SelecMix（selected mixup）（Hwang等，2022）在图像层面进行混合操作（MixUp）以实现图像增强；其余方法均使用了特征解耦的思想。从表2的结果可以看出，重加权、图像增强和特征解耦均可以有效地提升无偏测试集上的识别率。同时，对比这6种方法可以发现，ReBias（removing bias with bias）（Bahng等，2020）在Colored MNIST上取得了最佳的去偏见效果，而SelecMix（Hwang等，2022）在Corrupted CIFAR-10和bFFHQ上取得了最佳的去偏见效果，这证明了图像增强和特征解耦方法的有效性。虽然LfF（Nam等，2020）没有取得最好的结果，但实验证明这种简单的重加权方法在没有偏见标签的情况下也可以提升模型公平性。

表2　不同方法在Colored MNIST、Corrupted CIFAR-10和bFFHQ的无偏测试集上的识别率

Table 2　The accuracies of different methods on unbiased set of the Colored MNIST, Corrupted CIFAR-10 and bFFHQ datasets

/%

方法	Colored数据集	MNIST数据集			Corrupted CIFAR-10数据集				bFFHQ数据集
	ρ = 0.5%	ρ = 1%	ρ = 2%	ρ = 5%	ρ = 0.5%	ρ = 1%	ρ = 2%	ρ = 5%	ρ = 0.5%
Vanilla(He等，2016)	35.71±0.83	50.51±2.17	65.40±1.63	82.12±1.52	23.26±0.29	26.10±0.72	31.04±0.44	41.98±0.12	56.20±0.35
HEX（Wang等，2019b）	30.33±0.76	43.73±5.50	56.85±2.58	74.62±3.20	13.87±0.06	14.81±0.42	15.20±0.54	16.04±0.63	52.83±0.90
ReBias（Bahng等，2020）	**71.42±1.41**	**86.50±0.97**	**92.95±0.21**	**96.92±0.09**	22.13±0.23	26.05±0.10	32.00±0.81	44.00±0.66	56.80±1.56
EnD（Tartaglione等，2021）	56.98±4.85	73.83±2.09	82.28±1.08	89.26±0.27	22.54±0.65	26.20±0.39	32.99±0.33	44.90±0.37	56.53±0.61
LfF(Nam等，2020)	63.86±2.81	78.64±1.51	84.95±1.71	89.42±0.65	29.36±0.18	33.50±0.52	40.65±1.23	50.95±0.40	65.60±1.40
DFA（Lee等，2021）	67.37±1.61	80.20±1.86	85.61±0.76	89.86±0.80	30.04±0.66	33.80±1.83	42.10±1.04	49.23±0.63	61.60±1.97
SelecMix（Hwang等，2022）	70.00±0.52	82.80±0.71	87.16±0.62	91.57±0.20	**39.44±0.22**	**43.68±0.51**	**49.70±0.54**	**57.03±0.48**	**70.80±2.95**

注：加粗字体表示各列最优结果，ρ表示偏见冲突样本在训练集中的比例。

表3展示了10种去偏见方法在CelebA上的EOD结果，以衡量不同方法对性别和年龄偏见的去偏见效果。其中，MFD（MMD-based fair distillation）（Jung等，2021）、SupCon（Khosla等，2020）和FSCL（fair supervised contrastive loss）（Park等，2022）利用度量学习进行模型去偏；RNF（representation neutralization for fairness）（Du等，2021）采用特征增强使得分类器对偏见信息不敏感；DI（domain independent training）（Wang等，2022）基于后处理技术实现公平性；其余方法均使用特征解耦的思想消除偏见信息的负面影响。从表3的结果可以看出，度量学习、特征增强、特征解耦和后处理均能降低CelebA上的EOD，消除性别和年龄偏见。同时，对比这10种方法可以发现，基于度量学习的FSCL（Park等，2022）在5个子任务上取得最佳的去偏见效果，而基于特征解耦的AdvDebias（adversarial debias）（Wang等，2019c）也在部分任务上取得最佳的结果。度量学习在特征空间中约束不同样本间的距离，使得特征根据目标属性而非偏见属性进行聚类，因此，这种直接且强力的方法更易取得较好的效果。

表4和表5分别展示了使用BUPT-Globalface和BUPT-Balancedface进行训练并使用RFW进行测试时，6种去偏见方法在人脸识别任务上对种族偏见的去偏见效果。平均识别率反映了模型整体的准确率，而识别率方差则反映模型的公平性程度。其中，Re-weight（Ren等，2018）采用重加权实现去偏见；Adv（adversarial learning）（Alvi等，2018）和DebFace（debiasing adversarial network）（Gong等，2020）使用特征解耦消除种族对识别的影响；其余方法均采用模型自适应提升公平性。从表4和表5的结果可以看出，重加权、特征解耦和模型自适应均能提升人脸识别模型的公平性，降低模型在不同种族上的识别

表 3　不同方法在 CelebA 上的几率均等结果

Table 3　The equalized odds results of different methods on the CelebA dataset

/%

方法	T=吸引力 B=性别	T=金发 B=性别	T=大鼻子 B=性别	T=眼袋 B=性别	T=吸引力 B=年龄	T=金发 B=年龄	T=大鼻子 B=年龄	T=眼袋 B=年龄
Vanilla(He 等,2016)	26.24	40.82	23.93	15.00	20.52	4.05	18.49	12.70
AdvDebias(Wang 等,2019c)	11.56	33.44	15.96	–	**10.48**	**3.74**	7.12	–
GRL(Raff 和 Sylvester,2018)	24.90	–	14.00	6.70	14.70	–	10.00	5.90
LNL(Kim 等,2019a)	26.43	33.17	28.07	5.00	19.19	5.14	16.54	3.30
EnD(Tartaglione 等,2021)	24.64	33.73	22.04	–	21.57	3.91	17.65	–
MFD(Jung 等,2021)	20.17	38.84	28.86	8.70	22.00	5.16	16.12	5.20
DI(Wang 等,2020)	23.01	**7.76**	15.89	–	17.17	4.66	10.64	–
RNF(Du 等,2021)	40.15	24.01	23.58	–	22.42	5.23	14.36	–
SupCon(Khosla 等,2020)	30.50	–	20.70	20.80	21.70	–	16.90	10.80
FD-VAE(Park 等,2021)	15.10	–	11.20	5.70	14.80	–	6.70	6.20
FSCL(Park 等,2022)	**6.50**	–	**4.70**	**3.00**	12.40	–	**4.80**	**1.60**

注:加粗字体表示各列最优结果,T 表示目标属性,B 表示偏见属性,“–”表示未进行测试。

表 4　不同方法使用 BUPT-Globalface 训练在 RFW 数据集上的去偏见结果

Table 4　Debiasing results of different methods which are trained on BUPT-Globalface and evaluated on RFW

/%

方法	白人	印度人	东亚人	黑人	平均识别率(↑)	识别率方差(↓)
Vanilla(He 等,2016)	**97.37**	95.68	94.55	93.87	95.37	1.53
Re-weight(Ren 等,2018)	96.35	95.32	94.25	93.48	94.85	1.25
Adv(Alvi 等,2018)	96.63	95.27	94.17	93.70	94.94	1.30
RL-RBN(Wang 和 Deng,2020)	97.08	95.63	95.57	94.87	95.79	0.93
MBN(Wang 等,2022)	96.87	**96.20**	**95.63**	**95.00**	**95.93**	**0.80**

注:加粗字体表示各列最优结果,“↑”表示值越大越好,“↓”表示值越小越好。

表 5　不同方法使用 BUPT-Balancedface 训练在 RFW 数据集上的去偏见结果

Table 5　Debiasing results of different methods which are trained on BUPT-Balancedface and evaluated on RFW

/%

方法	白人	印度人	东亚人	黑人	平均识别率(↑)	识别率方差(↓)
Vanilla(He 等,2016)	96.18	94.67	93.72	93.98	94.64	1.11
DebFace(Gong 等,2020)	95.95	94.78	94.33	93.67	94.68	0.83
RL-RBN(Wang 和 Deng,2020)	**96.27**	94.68	94.82	95.00	95.19	0.73
GAC(Gong 等,2021)	96.20	94.98	**94.87**	94.77	95.21	**0.58**
MBN(Wang 等,2022)	96.25	**95.32**	94.85	**95.38**	**95.45**	**0.58**

注:加粗字体表示各列最优结果,“↑”表示值越大越好,“↓”表示值越小越好。

率方差。同时,对比这 6 种方法可以发现,相比于重加权和特征解耦,模型自适应在人脸识别的去偏见任务上表现更好。首先,特征解耦需要在准确率和公平性之间进行权衡。因此,为了保持令人满意的

准确率，公平性无法得到显著提高。其次，重加权方法只是通过对有色人种施加更大的权重，使网络在训练过程中更加关注有色人种。然而，在人脸识别这种细粒度的开集识别问题上，该方法对泛化性能的改进是非常有限的。模型自适应能够通过网络结构和超参数的自适应调整，满足不同群体的特殊需求，针对性地解决群体差异性带来的偏见问题，因此，可以达到更好的去偏见效果。

5 机遇与挑战

尽管学术界对图像识别公平性的研究已经取得了较大的进展，但仍存在各种悬而未决的挑战需要进一步关注：

1）数据集和评价指标仍有待完善。近年来，多个数据集被提出以研究偏见问题，如表1所示，然而，偏见属性仅限于颜色、背景、纹理、图像质量、种族、年龄和性别。现实世界中，偏见是多种多样的，因此，更加多样的公平性数据有待构造以研究更复杂的偏见问题。其次，现有的公平性数据集的构造方式仍存在差异，例如，在针对种族偏见的数据集中，种族的划分方式不同，RFW（Wang等，2019a）将种族划分为白人、印度人、东亚人和黑人4类，而Fairface（Kärkkäinen和Joo，2021）将种族划分为东亚人、南亚人、中东人、白人、拉美人、印度人和黑人7类。最后，评价指标尚未统一，不同去偏见算法选取不同的数据集和评价指标进行算法验证，且偏见对齐和偏见冲突样本的比例也不尽相同，这对比较和评估不同去偏见算法的优劣性造成困难。

2）针对未知偏见的公平性算法亟待解决。在真实世界中，“偏见是已知”的假设是不切实际的。获取潜在偏见的先验知识往往比较困难，需要了解潜在偏见的工人对偏见属性进行手动标记，这需要较强的专业知识，且费时费力。因此，在不依赖如此昂贵的偏见标签的情况下进行有效的公平性学习，是该领域所面临的现实性挑战。Seo等人（2022）和Jung等人（2022）利用聚类算法和辅助分类器为样本的偏见属性生成伪标签。Li等人（2022）利用发现器和分类器的迭代训练进行公平性学习，其中，发现器以最大化EOP为目标来发现多元且未知的偏见，分类器旨在消除发现器发现的偏见。Jeon等人（2022）观察到浅层特征和多样性更有利于模型公平性，因此，他们在不依赖于偏见标签的情况下，通过从网络各层捕获的分层特征和正交正则化来实现去偏见学习。

3）准确率和公平性的权衡困境有待突破。在图像识别的公平性研究中，除了要求消除模型在决策时对不同偏见属性的群体的决策差异外，同时还存在一项研究目标：尽可能保持模型在原始任务上的准确性，较好的模型准确性保证了模型的可用性。然而，不同领域公平性的研究已经证实，公平性和算法性能之间存在权衡困境（Zhao和Gordon，2022；Gajane和Pechenizkiy，2018；Fish等，2016），提高算法的公平性往往以降低性能为代价。由于公平性和性能都是不可或缺的，因此需要进行深入研究，以帮助人们更好地理解算法在两者之间的权衡机制，并设计方法能够在最大程度保持性能的情况下提升模型公平性。

4）针对细分任务的独特发展趋势开始呈现。在图像分类的公平性研究中，因果干预（Holland，1986）被引入，以建立正确的因果关系，避免模型依据偏见属性做出错误的判断和推理。例如，Zhang等人（2022）提出，在训练时利用代理特征捕捉偏见，并在推理阶段利用因果干预来消除代理特征从而去除偏见；Wang等人（2021）利用一个因果注意力模块使得模型能够重点关注目标区域进行推理。在人脸识别的公平性中，针对群体公平性的研究开始的较早。近期，一些工作（Sun等，2022）开始关注个体公平性，鼓励模型平等地对待每个个体，即具有相似图像分布的人应该得到相似的表现。

5）从图像公平性到视频公平性。视频数据的去偏见研究也是一个有意义且重要的领域。近年来，一些工作（Choi等，2019；Hazirbas等，2022）围绕该领域展开。例如，在UCF101（Soomro等，2012）和HMDB-51（human motion database）（Kuehne等，2011）等视频数据集中，动作识别任务易受到场景偏见的影响。为提升模型的公平性，Choi等人（2019）利用对抗学习和掩码消除偏见的负面作用；RESOUND（representation unbiased dataset）（Li等，2018）使用重加权方法来平衡数据集分布；Li等人（2023）提出一个简单的视频数据增强方法来对抗偏见。在属性识别和人脸识别领域，Hazirbas等人（2022）提出一个视频数据集，并在该数据集上测试了现有人脸属性识别和人脸识别算法针对性别、年

龄和肤色的偏见。

6 结　语

算法公平性是人工智能向善的重要主题之一，也是可信人工智能的重要组成部分，建立合理的模型以保证算法的无偏决策是加速推广图像识别落地的必要条件，具有理论意义和应用价值。本文对图像识别公平性领域自 2018 年兴起以来的去偏见主流算法：重加权（重采样）、图像增强、特征增强、特征解耦、度量学习、模型自适应和后处理进行综述。此外，还介绍了该领域的常用数据集、评价指标。最后对该领域的问题与挑战进行了总结并提出未来研究方向。希望通过本文让读者了解该领域工作前沿，以启发进而做出更有价值的公平性工作。

参考文献（References）

Ahn S，Kim S and Yun S Y. 2023. Mitigating dataset bias by using per-sample gradient//Proceedings of the 11th International Conference on Learning Representations. Kigali，Rwanda：OpenReview.net：1-14

Alvi M，Zisserman A and Nellåker C. 2018. Turning a blind eye：explicit removal of biases and variation from deep neural network embeddings//Proceedings of 2018 European Conference on Computer Vision. Munich，Germany：Springer：556-572［DOI：10.1007/978-3-030-11009-3_34］

Amini A，Soleimany A P，Schwarting W，Bhatia S N and Rus D. 2019. Uncovering and mitigating algorithmic bias through learned latent structure//Proceedings of 2019 AAAI/ACM Conference on AI，Ethics，and Society. Honolulu，USA：ACM：289-295［DOI：10.1145/3306618.3314243］

Bahng H，Chun S，Yun S，Choo J and Oh S J. 2020. Learning de-biased representations with biased representations//Proceedings of the 37th International Conference on Machine Learning. Vienna，Austria：PMLR：528-539

Bruveris M，Mortazavian P，Gietema J and Mahadevan M. 2020. Reducing geographic performance differentials for face recognition//Proceedings of 2020 IEEE Winter Applications of Computer Vision Workshops. Snowmass，USA：IEEE：98-106［DOI：10.1109/WACVW50321.2020.9096930］

Chapelle O，Haffner P and Vapnik V N. 1999. Support vector machines for histogram-based image classification. IEEE transactions on Neural Networks，10(5)：1055-1064［DOI：10.1109/72.788646］

Choi J，Gao C，Messou J C E and Huang J B. 2019. Why can't I dance in a mall？ Learning to mitigate scene bias in action recognition//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook，USA：Curran Associates Inc.：#77［DOI：10.5555/3454287.3454364］

Choi Y，Choi M，Kim M，Ha J W，Kim S and Choo J. 2018. Stargan：unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：8789-8797［DOI：10.1109/CVPR.2018.00916］

Chuang C Y and Mroueh Y. 2021. Fair mixup：fairness via interpolation//Proceedings of the 9th International Conference on Learning Representations. Virtual：OpenReview.net：1-11

Danziger S，Levav J and Avnaim-Pesso L. 2011. Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences of the United States of America，108(17)：6889-6892［DOI：10.1073/pnas.1018033108］

Deng J K，Guo J，Xue N N and Zafeiriou S. 2019. Arcface：additive angular margin loss for deep face recognition//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：4685-4694［DOI：10.1109/CVPR.2019.00482］

Deng W，Xing Y H，Li Y F，Li Z H and Wang G Y. 2020. Survey on fair machine learning. CAAI Transactions on Intelligent Systems，15(3)：578-586（邓蔚，邢钰晗，李逸凡，李振华，王国胤. 2020. 公平性机器学习研究综述. 智能系统学报，15(3)：578-586）［DOI：10.11992/tis.202007004］

Dhar P，Gleason J，Roy A，Castillo C D and Chellappa R. 2021. PASS：protected attribute suppression system for mitigating bias in face recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal，Canada：IEEE：15067-15076［DOI：10.1109/ICCV48922.2021.01481］

Du M N，Mukherjee S，Wang G C，Tang R X，Awadallah A and Hu X. 2021. Fairness via representation neutralization//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual：NeurIPS：12091-12103

Esteva A，Kuprel B，Novoa R A，Ko J，Swetter S M，Blau H M and Thrun S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature，542(7639)：115-118［DOI：10.1038/nature21056］

Fish B，Kun J and Lelkes Á D. 2016. A confidence-based approach for balancing fairness and accuracy//Proceedings of 2016 SIAM International Conference on Data Mining. Miami，USA：SIAM：144-152［DOI：10.1137/1.9781611974348.17］

Gajane P and Pechenizkiy M. 2018. On formalizing fairness in prediction with machine learning［EB/OL］.［2023-04-03］. https://arxiv.org/pdf/1710.03184.pdf

Ganin Y and Lempitsky V. 2015. Unsupervised domain adaptation by backpropagation//Proceedings of the 32nd International Conference on Machine Learning. Lille，France：PMLR：1180-1189

Ge J C，Deng W H，Wang M and Hu J N. 2020. FGAN：fan-shaped

GAN for racial transformation//Proceedings of 2020 IEEE International Joint Conference on Biometrics. Houston, USA: IEEE: 1-7 [DOI: 10.1109/IJCB48548.2020.9304901]

Georgopoulos M, Oldfield J, Nicolaou M A, Panagakis Y and Pantic M. 2021. Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. International Journal of Computer Vision, 129 (7): 2288-2307 [DOI: 10.1007/s11263-021-01448-w]

Gong S X, Liu X M and Jain A K. 2020. Jointly de-biasing face recognition and demographic attribute estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 330-347 [DOI: 10.1007/978-3-030-58526-6_20]

Gong S X, Liu X M and Jain A K. 2021. Mitigating face recognition bias via group adaptive classifier//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3413-3423 [DOI: 10.1109/CVPR46437.2021.00342]

Guo Y D, Zhang L, Hu Y X, He X D and Guo J F. 2016. MS-celeb-1M: a dataset and benchmark for large-scale face recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 87-102 [DOI: 10.1007/978-3-319-46487-9_6]

Hardt M, Price E and Srebro N. 2016. Equality of opportunity in supervised learning//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS: 3323-3331

Hazirbas C, Bitton J, Dolhansky B, Pan J, Gordo A and Ferrer C C. 2022. Towards measuring fairness in AI: the casual conversations dataset. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(3): 324-332 [DOI: 10.1109/TBIOM.2021.3132237]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He Z L, Zuo W M, Kan M M, Shan S G and Chen X L. 2019. AttGAN: facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11): 5464-5478 [DOI: 10.1109/TIP.2019.2916751]

Hendrycks D and Dietterich T G. 2019. Benchmarking neural network robustness to common corruptions and perturbations//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net: 1-10

Holland P W. 1986. Statistics and causal inference. Journal of the American statistical Association, 81(396): 945-960 [DOI: 10.1080/01621459.1986.10478354]

Hong Y and Yang E. 2021. Unbiased classification through bias-contrastive and bias-balanced learning//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: NeurIPS: 26449-26461

Huang R, Geng A and Li Y X. 2021. On the importance of gradients for detecting distributional shifts in the wild//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: NeurIPS: 677-689

Huang X and Belongie S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1510-1519 [DOI: 10.1109/ICCV.2017.167]

Hwang I, Lee S, Kwak Y, Oh S J, Teney D, Kim J H and Zhang B T. 2022. SelecMix: debiased learning by contradicting-pair sampling//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS

Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B and Madry A. 2019. Adversarial examples are not bugs, they are features//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: 125-136

Iurada L, Bucci S, Hospedales T M and Tommasi T. 2023. Fairness meets cross-domain learning: a new perspective on models and metrics [EB/OL]. [2023-04-03]. https://arxiv.org/pdf/2303.14411.pdf

Jeon M, Kim D, Lee W, Kang M and Lee J. 2022. A conservative approach for unbiased learning on unknown biases//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16731-16739 [DOI: 10.1109/CVPR52688.2022.01625]

Jung S, Chun S and Moon T. 2022. Learning fair classifiers with partially annotated group labels//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10338-10347 [DOI: 10.1109/CVPR52688.2022.01010]

Jung S, Lee D, Park T and Moon T. 2021. Fair feature distillation for visual recognition//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 12110-12119 [DOI: 10.1109/CVPR46437.2021.01194]

Kärkkäinen K and Joo J. 2021. FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 1547-1557 [DOI: 10.1109/WACV48630.2021.00159]

Karras T, Laine S and Aila T. 2019. A style-based generator architecture for generative adversarial networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4396-4405 [DOI: 10.1109/CVPR.2019.00453]

Khosla P, Teterwak P, Wang C, Sarna A, Tian Y L, Isola P, Maschinot A, Liu C and Krishnan D. 2020. Supervised contrastive learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: #1567 [DOI: 10.5555/3495724.3497291]

Kim B, Kim H, Kim K, Kim S and Kim J. 2019a. Learning not to

learn: training deep neural networks with biased data//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9004-9012 [DOI: 10.1109/CVPR.2019.00922]

Kim E, Lee J and Choo J. 2021. BiaSwap: removing dataset bias with bias-tailored swapping augmentation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14972-14981 [DOI: 10.1109/ICCV48922.2021.01472]

Kim M P, Ghorbani A and Zou J. 2019b. Multiaccuracy: black-box post-processing for fairness in classification//Proceedings of 2019 AAAI/ACM Conference on AI, Ethics, and Society. Honolulu, USA: ACM: 247-254 [DOI: 10.1145/3306618.3314287]

Kim N, Hwang S, Ahn S, Park J and Kwak S. 2022. Learning debiased classifier with biased committee//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: NeurIPS: 1-13

Klare B F, Burge M J, Klontz J C, Bruegge R W V and Jain A K. 2012. Face recognition performance: role of demographic information. IEEE Transactions on Information Forensics and Security, 7(6): 1789-1801 [DOI: 10.1109/TIFS.2012.2214212]

Krizhevsky A. 2009. Learning Multiple Layers of Features from Tiny Images. Toronto: University of Toronto

Krizhevsky A, Sutskever I and Hinton G E. 2017. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6): 84-90 [DOI: 10.1145/3065386]

Kuehne H, Jhuang H, Garrote E, Poggio T and Serre T. 2011. HMDB: a large video database for human motion recognition//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 2556-2563 [DOI: 10.1109/ICCV.2011.6126543]

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

Lee J, Kim E, Lee J, Lee J and Choo J. 2021. Learning debiased representation via disentangled feature augmentation//Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual: NeurIPS: 25123-25133

Li H X, Liu Y, Zhang H W and Li B Y. 2023. Mitigating and evaluating static bias of action representations in the background and the foreground [EB/OL]. [2023-04-03]. https://arxiv.org/pdf/2211.12883.pdf

Li Y and Vasconcelos N. 2019. Repair: removing representation bias by dataset resampling//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9564-9573 [DOI: 10.1109/CVPR.2019.00980]

Li Y W, Li Y and Vasconcelos N. 2018 RESOUND: towards action recognition without representation bias//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springe: 520-535 [DOI: 10.1007/978-3-030-01231-1_32]

Li Z H, Hoogs A and Xu C L. 2022. Discover and mitigate unknown biases with debiasing alternate networks//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 270-288 [DOI: 10.1007/978-3-031-19778-9_16]

Liu W Y, Shen C Y, Wang X F, Jin B, Lu X J, Wang X L, Zha H Y and He J F. 2021. Survey on fairness in trustworthy machine learning. Journal of Software, 32(5): 1404-1426 (刘文炎, 沈楚云, 王祥丰, 金博, 卢兴见, 王晓玲, 查宏远, 何积丰. 2021. 可信机器学习的公平性综述. 软件学报, 32(5): 1404-1426) [DOI: 10.13328/j.cnki.jos.006214]

Liu Z W, Luo P, Wang X G and Tang X O. 2015. Deep learning face attributes in the wild//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3730-3738 [DOI: 10.1109/ICCV.2015.425]

Mehrabi N, Morstatter F, Saxena N, Lerman K and Galstyan A. 2021. A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6): #115 [DOI: 10.1145/3457607]

Nam J H, Cha H, Ahn S, Lee J and Shin J. 2020. Learning from failure: De-biasing classifier from biased classifier//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: NeurIPS: 20673-20684

Nuriel O, Benaim S and Wolf L. 2021. Permuted AdaIN: reducing the bias towards global statistics in image classification//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9477-9486 [DOI: 10.1109/CVPR46437.2021.00936]

Park S, Hwang S, Kim D and Byun H. 2021. Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual: AAAI: 2403-2411 [DOI: 10.1609/aaai.v35i3.16341]

Park S, Lee J, Lee P, Hwang S, Kim D and Byun H. 2022. Fair contrastive learning for facial attribute classification//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10379-10388 [DOI: 10.1109/CVPR52688.2022.01014]

Raff E and Sylvester J. 2018. Gradient reversal against discrimination: a fair neural network learning approach//Proceedings of 2018 IEEE 5th International Conference on Data Science and Advanced Analytics. Turin, Italy: IEEE: 189-198 [DOI: 10.1109/DSAA.2018.00029]

Ragonesi R, Volpi R, Cavazza J and Murino V. 2021. Learning unbiased representations via mutual information backpropagation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 2723-2732 [DOI: 10.1109/CVPRW53098.2021.00307]

Ramaswamy V V, Kim S S Y and Russakovsky O. 2021. Fair attribute classification through latent space de-biasing//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion. Nashville, USA: IEEE: 9297-9306 [DOI: 10.1109/CVPR46437.2021.00918]

Ren M Y, Zeng W Y, Yang B and Urtasun R. 2018. Learning to reweight examples for robust deep learning//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 4334-4343

Rothe R, Timofte R and Van Gool L. 2018. Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision, 126(2/4): 144-157 [DOI: 10.1007/s11263-016-0940-3]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Sánchez J, Perronnin F, Mensink T and Verbeek J. 2013. Image classification with the fisher vector: theory and practice. International Journal of Computer Vision, 105(3): 222-245 [DOI: 10.1007/s11263-013-0636-x]

Sarhan M H, Navab N, Eslami A and Albarqouni S. 2020. Fairness by learning orthogonal disentangled representations//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 746-761 [DOI: 10.1007/978-3-030-58526-6_44]

Saxena N A, Huang K R, DeFilippis E, Radanovic G, Parkes D C and Liu Y. 2019. How do fairness definitions fare?: examining public attitudes towards algorithmic definitions of fairness//Proceedings of 2019 AAAI/ACM Conference on AI, Ethics, and Society. Honolulu, USA: ACM: 99-106 [DOI: 10.1145/3306618.3314248]

Seo S, Lee J Y and Han B. 2022. Unsupervised learning of debiased representations with pseudo-attributes//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 16721-16730 [DOI: 10.1109/CVPR52688.2022.01624]

Seyyed-Kalantari L, Zhang H R, McDermott M B A, Chen I Y and Ghassemi M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine, 27(12): 2176-2182 [DOI: 10.1038/s41591-021-01595-0]

Shrestha R, Kafle K and Kanan C. 2022a. An investigation of critical issues in bias mitigation techniques//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2512-2523 [DOI: 10.1109/WACV51458.2022.00257]

Shrestha R, Kafle K and Kanan C. 2022b. Occamnets: mitigating dataset bias by favoring simpler hypotheses//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 702-721 [DOI: 10.1007/978-3-031-20044-1_40]

Singh R, Majumdar P, Mittal S and Vatsa M. 2022. Anatomizing bias in facial analysis//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtual: AAAI: 12351-12358 [DOI: 10.1609/aaai.v36i11.21500]

Soomro K, Zamir A R and Shah M. 2012. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. [2023-04-03]. https://arxiv.org/pdf/1212.0402.pdf

Sun T, Gaut A, Tang S, Huang Y X, ElSherief M, Zhao J Y, Mirza D, Belding E, Chang K W and Wang W Y. 2019. Mitigating gender bias in natural language processing: literature review//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL: 1630-1640 [DOI: 10.18653/v1/P19-1159]

Sun Y F, Li Y and Cui Z. 2022. NFW: towards national and individual fairness in face recognition//Proceedings of the 6th Asian Conference on Pattern Recognition. Jeju Island, Korea (South): Springer: 540-553 [DOI: 10.1007/978-3-031-02375-0_40]

Tartaglione E, Barbano C A and Grangetto M. 2021. EnD: entangling and disentangling deep representations for bias correction//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 13503-13512 [DOI: 10.1109/CVPR46437.2021.01330]

Terhörst P, Tran M L, Damer N, Kirchbuchner F and Kuijper A. 2020a. Comparison-level mitigation of ethnic bias in face recognition//Proceedings of the 8th International Workshop on Biometrics and Forensics. Porto, Portugal: IEEE: 1-6 [DOI: 10.1109/IWBF49977.2020.9107956]

Terhörst P, Kolf J N, Damer N, Kirchbuchner F and Kuijper A. 2020b. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. Pattern Recognition Letters, 140: 332-338 [DOI: 10.1016/j.patrec.2020.11.007]

Tzeng E, Hoffman J, Saenko K and Darrell T. 2017. Adversarial discriminative domain adaptation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2962-2971 [DOI: 10.1109/CVPR.2017.316]

Wang H, Wang Y T, Zhou Z, Ji X, Gong D H, Zhou J C, Li Z F and Liu W. 2018. CosFace: large margin cosine loss for deep face recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5265-5274 [DOI: 10.1109/CVPR.2018.00552]

Wang H H, He Z X, Lipton Z C and Xing E P. 2019b. Learning robust representations by projecting superficial statistics out//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA: OpenReview.net

Wang M and Deng W H. 2020. Mitigating bias in face recognition using skewness-aware reinforcement learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9319-9328 [DOI: 10.1109/CVPR42600.2020.00934]

Wang M, Deng W H, Hu J N, Tao X Q and Huang Y H. 2019a. Racial faces in the wild: Reducing racial bias by information maximization

adaptation network//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 692-702 [DOI: 10.1109/ICCV.2019.00078]

Wang M, Zhang Y B and Deng W H. 2022. Meta balanced network for fair face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44 (11): 8433-8448 [DOI: 10.1109/TPAMI.2021.3103191]

Wang T, Zhou C, Sun Q R and Zhang H W. 2021. Causal attention for unbiased visual recognition//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 3071-3080 [DOI: 10.1109/ICCV48922.2021.00308]

Wang T L, Zhao J Y, Yatskar M, Chang K W and Ordonez V. 2019c. Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5309-5318 [DOI: 10.1109/ICCV.2019.00541]

Wang Y F, Ma W Z, Zhang M, Liu Y Q and Ma S P. 2023. A survey on the fairness of recommender systems. ACM Transactions on Information Systems, 41(3): #52 [DOI: 10.1145/3547333]

Wang Z Y, Qinami K, Karakozis I C, Genova K, Nair P, Hata K and Russakovsky O. 2020. Towards fairness in visual recognition: effective strategies for bias mitigation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8916-8925 [DOI: 10.1109/CVPR42600.2020.00894]

Yao B P, Jiang X Y, Khosla A, Lin A L, Guibas L and Li F F. 2011. Human action recognition by learning bases of action attributes and parts//Proceedings of 2011 International Conference on Computer Vision. Barcelona, Spain: IEEE: 1331-1338 [DOI: 10.1109/ICCV.2011.6126386]

Yatskar M, Zettlemoyer L and Farhadi A. 2016. Situation recognition: visual semantic role labeling for image understanding//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 5534-5542 [DOI: 10.1109/CVPR.2016.597]

Yucer S, Akçay S, Al-Moubayed N and Breckon T P. 2020. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 83-92 [DOI: 10.1109/CVPRW50498.2020.00017]

Zhang F D, Kuang K, Chen L, Liu Y X, Wu C and Xiao J. 2023. Fairness-aware contrastive learning with partially annotated sensitive attributes//Proceedings of the 11th International Conference on Learning Representations. Kigali, Rwanda: OpenReview.net

Zhang Y and Sang J T. 2020. Towards accuracy-fairness paradox: adversarial example-based data augmentation for visual debiasing//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 4346-4354 [DOI: 10.1145/3394171.3413772]

Zhang Y, Sang J T and Wang J Y. 2022. Fair visual recognition via intervention with proxy features [EB/OL]. [2023-04-03]. https://arxiv.org/pdf/2211.01253.pdf

Zhang Z F, Song Y and Qi H R. 2017. Age progression/regression by conditional adversarial autoencoder//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 4352-4360 [DOI: 10.1109/CVPR.2017.463]

Zhao B W, Chen C, Wang Q W, He A F and Xia S T. 2021. Combating unknown bias with effective bias-conflicting scoring and gradient alignment [EB/OL]. [2023-04-03]. https://arxiv.org/pdf/2111.13108.pdf

Zhao H and Gordon G J. 2022. Inherent tradeoffs in learning fair representations. The Journal of Machine Learning Research, 23(1): #57 [DOI: 10.5555/3586589.3586646]

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu W, Zheng H T, Liao H F, Li W J and Luo J B. 2021. Learning bias-invariant representation by cross-sample mutual information minimization//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 14982-14992 [DOI: 10.1109/ICCV48922.2021.01473]

Zou J and Schiebinger L. 2018. AI can be sexist and racist—it's time to make it fair. Nature, 559 (7714): 324-326 [DOI: 10.1038/d41586-018-05707-8]

作者简介

王玫，女，副教授，主要研究方向为计算机视觉、可信图像识别和迁移学习。E-mail: wangmei1@bn.edu.cn

邓伟洪，通信作者，男，教授，博士生导师，主要研究方向为模式识别与计算机视觉、人脸识别、表情识别、行人再识别和细粒度图像识别。E-mail: whdeng@bupt.edu.cn

苏森，男，教授，博士生导师，主要研究方向为数据隐私和云计算。E-mail: susen@bupt.edu.cn