Neural Computing and Applications (2022) 34:12875–12893
https://doi.org/10.1007/s00521-022-07136-1

ORIGINAL ARTICLE

Image fairness in deep learning: problems, models, and challenges

Huan Tian[1] · Tianqing Zhu[1] · Wei Liu[1] · Wanlei Zhou[2]

Received: 16 August 2021 / Accepted: 22 February 2022 / Published online: 25 March 2022
© The Author(s) 2022

Abstract

In recent years, it has been revealed that machine learning models can produce discriminatory predictions. Hence, fairness protection has come to play a pivotal role in machine learning. In the past, most studies on fairness protection have used traditional machine learning methods to enforce fairness. However, these studies focus on low dimensional inputs, such as numerical inputs, whereas more recent deep learning technologies have encouraged fairness protection with image inputs through deep model methods. These approaches involve various object functions and structural designs that break the spurious correlations between targets and sensitive features. With these connections broken, we are left with fairer predictions. To better understand the proposed methods and encourage further development in the field, this paper summarizes fairness protection methods in terms of three aspects: the problem settings, the models, and the challenges. Through this survey, we hope to reveal research trends in the field, discover the fundamentals of enforcing fairness, and summarize the main challenges to producing fairer models.

Keywords Image fairness protection · Deep learning · Fair representations

1 Introduction

Protecting fairness in a machine learning model means measuring and eliminating discrimination in the model and ensuring the applications built around the model are trustworthy. The aim is to prevent the model from making significantly different predictions for different sub-groups, where each subgroup is divided by a “sensitive feature”, such as race, gender, or age. The phenomenon of unfairness has been observed frequently across the machine learning field. For example, software used to support recruitment and hiring decisions has been found to be discriminatory [1]. And there have been severe problems with gender bias in the Amazon AI curriculum selection [2]. In fact, machine learning models are so widely deployed in our society that, without fairness protection, we may find the impacts of discrimination to be catastrophic [3–8].

Fortunately, studies on fairness have a long history [9], so there is much documented evidence of not only the biases that lead to unfair predictions [10–12] but also a plethora of approaches to overcome those biases. Suggested strategies include data sampling, re-weighting, and modification methods to enforce equal predictions across subgroups, along with different fairness metrics to measure the differences in predictions.

Most previous articles on fairness have focused on numerical or tablet inputs. However, there are a growing number of studies dealing with fairness protection for image inputs. Rapid developments in deep learning have seen various image datasets emerge, such as ImageNet [13] and KITTI [14], and, as in the past, unfairness and discrimination have again been observed with image inputs and deep model deployments. Buolamwini and Gebru [15] find that commercial face recognition systems suffer significant prediction gaps across populations, while Brandao [16] find age and gender biases among pedestrian detection algorithms. Unfair predictions jeopardize model

✉ Tianqing Zhu
tianqing.zhu@uts.edu.au

Huan Tian
huan.tian@student.uts.edu.au

Wei Liu
wei.liu@uts.edu.au

Wanlei Zhou
wlzhou@cityu.edu.mo

[1] School of Computer Science, University of Technology Sydney, Sydney, Australia

[2] City University of Macau, Macau, China

performance for minorities and lead to negative social impacts. Moreover, this phenomenon has also led to concerns over the training of deep models in that they tend to learn short-cut features that are irrelevant to the learning targets. Thus, the central issue in fairness protection for images is to break this short-cut learning so as to avoid unfair predictions.

It is natural to adopt previous methods for fair protection with image inputs. However, traditional value modification methods cannot hold. As discussed in [17, 18], an image feature, such as race, cannot be modified from one attribute to another, and enforcing fairness with other methods is inefficient compared to deep model methods [19]. Intuitively, simply balancing the training dataset should resolve the problem. However, as indicated in [20, 21], constructing a balanced dataset in all its attributes can be very challenging. What's more, even with balanced datasets, bias in the trained model cannot be eliminated completely [22].

The more recent deep learning methods try to enforce fairness protection with images through additional constraints, by removing sensitive features, and/or by learning fair representations. These strategies are often applied during training with the overarching objective of minimizing prediction gaps across the subgroups. The main challenge with this work is to remove any spurious bias that favors one subgroup over another. Part of this involves deriving invariant features during the learning process that will generalize well across domains. Broadly speaking, deep model studies in image fairness encourage a deeper understanding of the dynamic learning procedures of the model.

To better understand these studies and to encourage further development in the field, we, with this paper, have summarized the deep learning based fairness protection methods for images. Beyond highlighting the differences between image and numerical inputs, we also aim to reveal research trends in the field; summarize the methods and expose the fundamentals that bind these approaches; and discuss the main challenges researchers are currently facing in ensuring better protection.

Fairness concerns a range of fields. For example, different biases, such as historical bias, measurement bias, and evaluation bias, lead to different types of unfair predictions [10]. Further, fairness is also a subject that concerns social issues and impacts [11, 23]. Grgic-Hlaca et al. [24], for example, discuss definitions of fairness among different groups of people, while Mehrabi et al. [10] regard social historical reasons as sources of unfair bias. Hence, to narrow down the scope of this survey, we have concentrated on research in the field, considering only studies on representation bias—also known as data imbalance bias—which is the most commonly studied bias.

Additionally, there are some traditional image processing methods that can enforce fairness that have not been covered, both in the pre-processing phase, such as data sampling and re-weighting, or in the post-processing phase, such as parameter post tuning. As been discussed, these methods are inefficient compared with deep models and we do not include these methods to avoid repeating work already undertaken in other surveys. Our focus remains fixed on deep model methods. We also note that deep model studies from other fields, such as domain adaptation, could also be adopted for enforcing image fairness. These methods were included if they considered sensitive features, such as gender and race.

There have been several other surveys on fairness protection, each with their own distinct focus. For example, Mitchell et al. [3] concentrate on summarizing fair notions and metrics. Mehrabi et al. [10] focus on the sources of discrimination source. Quy et al. [25] examine the underlying relations in the attributes of fairness protection datasets. Caton and Haas [26] provide an overview of fairness protection from metrics to approaches to dilemmas in the machine learning area. As for fairness protection with deep models, Malik and Singh [27] discuss general deep learning technology, offering an introduction to unfair interpretation. Du et al. [28] present deep methods in terms of the bias found in inputs and representations, while Shi [29] looks at issues of unfairness in deep federated learning methods. Our work provides a thorough summary of image fairness protection with deep models. We present a comprehensive view of the problems, models, and challenges associated in this area. Unlike other surveys, we have analyzed the research trends in fairness protection for deep image models. We have also pinpointed three fundamental challenges to better fairness and discussed solutions drawn from other fields. Table 1 lists the different scopes of each of the fairness surveys.

In summary, this work contributes the following additions to the literature:

- We have highlighted the difference between numerical and image inputs, summarizing the different problem settings for image inputs with deep models;
- Research trends in the field are extracted and outlined;
- The methods reviewed are classified into four approaches and compared against different fairness characteristics; and
- Three fundamental challenges for better fairness protection have been identified with potential solutions introduced from other fields.

Table 1 A comparison of different fairness protection surveys

Surveys	Main scope and content
Mitchell et al. [3]	Focuses on fair notions and metrics
Mehrabi et al. [10]	Summarizes different sources of discrimination
Quy et al. [25]	Examines underlying relations across attributes in datasets
Haas [26]	Provides reviews of fairness protection in the machine learning area
Malik and Singh [27]	An introduction to general deep learning and unfair interpretation
Du et al [28]	Summarizes deep methods in terms of biased inputs and representations
Shi [29]	Focuses on unfairness issues in deep federated learning methods
Our	Summarizes image fairness protection in deep models in terms of problems, models, and challenges

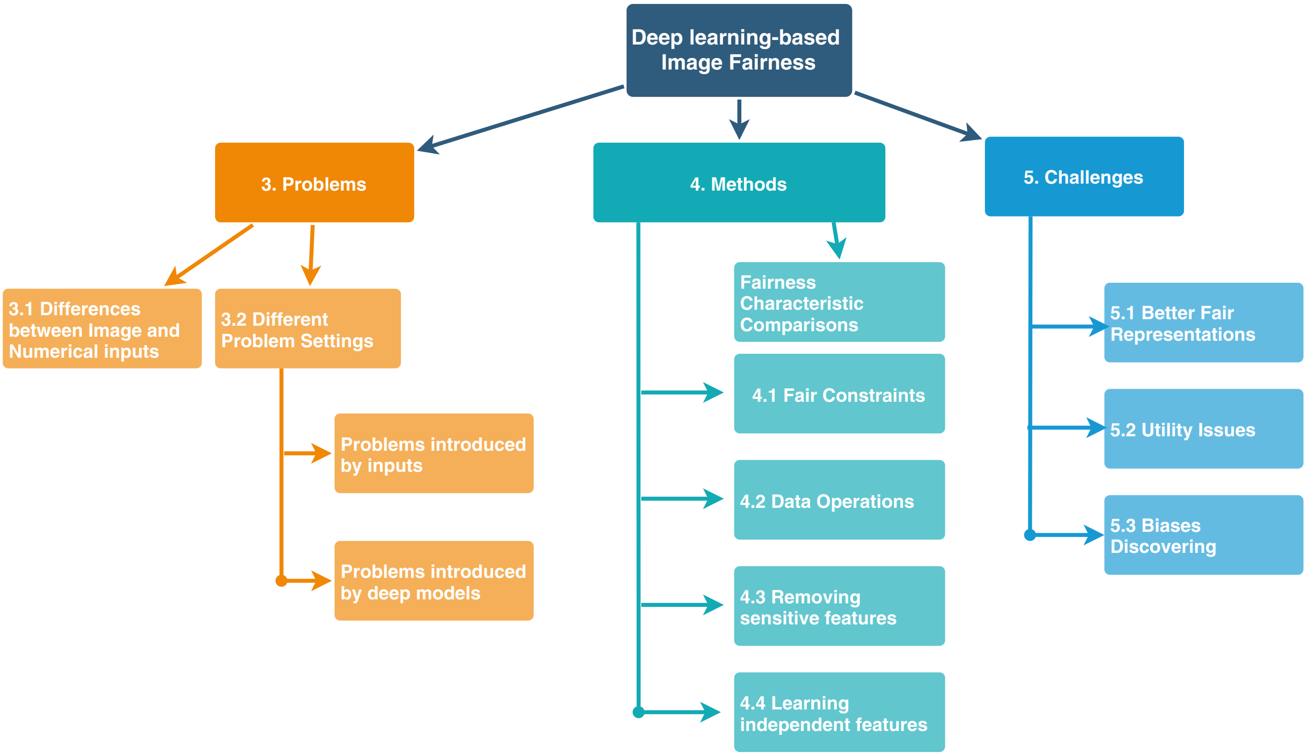

Deep learning-based Image Fairness
3. Problems
4. Methods
5. Challenges
3.1 Differences between Image and Numerical inputs
3.2 Different Problem Settings
Problems introduced by inputs
Problems introduced by deep models
Fairness Characteristic Comparisons
4.1 Fair Constraints
4.2 Data Operations
4.3 Removing sensitive features
4.4 Learning independent features
5.1 Better Fair Representations
5.2 Utility Issues
5.3 Biases Discovering

Fig. 1 The main structure of this survey: introducing image fairness protections with deep model methods from problems, models, and challenges

Figure 1 shows the main structure of this survey. The survey begins with background information on fairness protection in Sect. 2. Research trends with different problem settings are then discussed in Sect. 3. The deep model methods for protecting fairness are introduced in Sect. 4. Section 5 re-iterates the methods in terms of three challenges. Additionally, as fairness and privacy are closely related, we discuss this issue in Sect. 6. Future directions and conclusions are presented in Sect. 6.4.

2 Background and preliminary

Before jumping into image fairness methods, we need to introduce the background and preliminaries of image fairness. This includes the relevant definitions, notions and measurements, datasets, and the methods associated with protecting fairness with images. We will also introduce the most widely adopted and fundamental deep models in image fairness protection studies.

2.1 Definitions

The most common case considered in fairness studies is a binary classification problem with data $X \in R^n$, targets $Y \in \{0, 1\}$ and sensitive attributes $S \in \{0, 1\}$. The aim of enforcing fair predictions is to learn a model $f : X \rightarrow Y$ whose predictions $\hat{Y} \in \{0, 1\}$ are maximally close to Y while being fair for S under biased representations of training data.

Figure 2 illustrates a biased representation problem, also referred to as imbalance bias. In the image, the circles and squares are classification targets y_0, y_1, while the colors are sensitive features s_0, s_1. A well-trained and fair model should classify the shapes independently of the colors. However, the classifier favors blue samples over green ones due to short-cut learning issues and an over-representation of blue training samples. This spurious learning of features leads to unfair predictions. The most considered sensitive features in fairness studies are age, race, and gender.

2.2 Fairness notions and measurements

Aside from the social, ethical, or philosophical debates on defining fairness [30], studies usually consider fairness notions in three respects: (1) individually, where similar examples should be treated similarly [31]; (2) causally, where sensitive features should be independent of the target predictions [32–34]; and (3) as a group where different subgroups of sensitive features should process similar outputs [31, 33, 35]. Group fairness notions can also be regarded as statistical fairness notions since they compare statistics, such as accuracy or false positive rates. For example, demographic parity (DP) [31, 33] compares positive predictions results across subgroups:

$$P_{s_0}\{\hat{Y} = 1\} = P_{s_1}\{\hat{Y} = 1\},$$

Equalized opportunity (EOP) [35] compares equal true positive rates across subgroups:

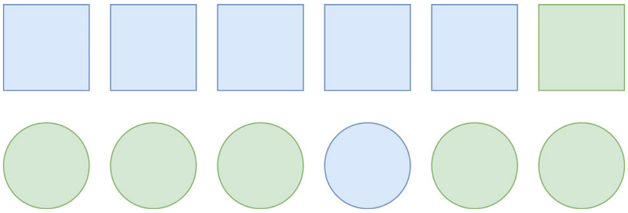


Fig. 2 Representation bias (imbalanced bias)

$$P_{s_0}\{\hat{Y} = 1|Y = 1\} = P_{s_1}\{\hat{Y} = 1|Y = 1\},$$

And equalized odds (EOD) [35] compares both true and false positive rates across subgroups:

$$P_{s_0}\{\hat{Y} = 1|Y = i\} = P_{s_1}\{\hat{Y} = 1|Y = i\}, i = 0, 1,$$

Different fair notions need to be considered in different scenarios. For example, individual fairness requires that similar samples get similar treatments, which is suitable for general fair tasks. However, similarity should be defined for a particular task. This is generally challenging to define [36]. Causal fairness notions should be adopted when causal graphs are considered. One example is counterfactual fairness [33], which requires similar predictions between samples and their counterfactual counterparts. However, specifying either the counterparts or the causal relationships between datasets is challenging. Group fairness is most widely adopted since it involves explicit constraints.

Fair measurements adopt fair notions to quantify fairness results. For example, disparate impact (DI) adopts demographic parity to ensure fair measurements:

$$\mathrm{DI} = \frac{P_{s_0}\{\hat{Y} = 1\}}{P_{s_1}\{\hat{Y} = 1\}},$$

As has been illustrated, disparate impact DI lies in the range $[0, \infty)$, where 1 denotes perfect demographic parity. A DI of <1 indicates that the classifier favors privileged groups and $\mathrm{DI} > 1$ means the opposite. Alternatively, subtractions with demographic parity can also be considered as measurements:

$$\Delta_{\mathrm{DP}} = |P_{s_0}\{\hat{Y} = 1\} - P_{s_1}\{\hat{Y} = 1\}|,$$

The metrics true positive rate balance (TPRB) and true negative rate balance (TNRB) have also been adopted alongside the notions of equalized opportunity and equalized odds:

$$\mathrm{TPRB} = P_{s_0}\{\hat{Y} = 1|Y = 1\} - P_{s_0}\{\hat{Y} = 1|Y = 1\},$$
$$\mathrm{TNRB} = P_{s_0}\{\hat{Y} = 1|Y = 0\} - P_{s_0}\{\hat{Y} = 1|Y = 0\},$$
$$\Delta_{\mathrm{EP}} = \frac{1}{2}\mathrm{TPRB} + \frac{1}{2}\mathrm{TNRB},$$

Note that these are by no means the only fairness notions and measurements. In their survey, Mehrabi et al. [10] introduce the most commonly used fair notions, while Islam et al. [37] summarize more than 20 different notions and measurements. A more detailed comparison of these different notions and measures and when they are used follows in Sect. 4.

2.3 Image fairness protection methods

Methods in fairness protection studies are generally categorized into three groups: (1) pre-processing methods such as sampling and re-weighting [38–40], where manipulations of training data are conducted before training; (2) in-processing methods [32, 41, 42], where the methods adopt additional losses for different fair metrics during the learning of the models; and (3) post-processing methods [43–45], where the methods tune the prediction results or adjust the decision boundaries to reduce unfair predictions.

Most surveys generally classify deep model methods as in-processing methods without much further discussion. However, since our focus is on deep models, we have divided the methods into four different approaches: fair constraint methods, data operation methods, methods that remove sensitive features, and methods that learn independent features. Each group is introduced in more detail in Sect. 4.

2.4 Datasets

Many of the studies on fairness rely on numerical datasets, such as Adult [46] and COMPAS [47]. The Adult dataset provides 48,842 records of people's salaries including the attributes of race and gender. The COMPAS dataset scores a criminal defendant's likelihood of re-offending (recidivism) with annotated attributions of over 10,000 samples.

The most widely-adopted image datasets for fairness studies are UTKFace [48] and CelebA [49]. UTKFace dataset consists of over 20,000 face images with annotations including age and ethnicity. The CelebA dataset provides more than 200K celebrity images with 40 attribute annotations. Generally speaking, age, gender, and race are commonly considered in image fairness studies.

Other datasets that have featured in papers on image fairness include the German credit dataset [46], the Diversity in Faces dataset [50], and the Yale B face dataset [51]. The Mehrabi et al. survey [10] also discusses some additional datasets. Table 2 provides a summary.

2.5 Deep learning and fairness

Most deep models for image fairness protection models are adopting generative models [52, 53] with adversarial designs that generate latent features or synthetic data from the original training data [54, 55]. The most common implementations are variational autoencoders (VAEs) [56, 57] based models or generative adversarial networks (GANs) [58] based models. VAE models require distribution similarities between the generated data and the original training data through similarity constraints. GAN models generate data with additional adversarial model designs. To enforce fairness, deep models generate synthetic images with generative models and remove sensitive features with adversarial designs.

3 Problem settings for image fairness with deep model methods

This section begins with an analysis of the difference between numerical and image inputs in terms of fairness. This is followed by the various problem settings and contexts in which research on these methods has been presented. Our discussions cover both data inputs and deep learning technologies, and through these discussions, research trends in the field emerge.

3.1 Input differences with images

In terms of numerical inputs, sensitive features, such as gender, age, or race, are generally represented as discrete values or as a binary variable $\{0, 1\}$. Because it is possible to modify these sensitive features, fairness-aware methods such as data modification [59] or data generation [60] have emerged as a solution. However, with image data, feature disentanglement in high dimensional domains makes representing features with explicit values generally impossible. Figure 3 illustrates the difference. As described by Mo et al. [21], due to this disentanglement, general datasets suffer background bias and content bias, and constructing a balanced dataset that fairly considers all the different attributes is a challenging task [20]. What's more, Xu et al. [22] observed that, even with balanced datasets containing equal training samples from a mixture of races, the trained model still suffered from unfair predictions. Another difference is that iterating sensitive feature is impossible because, unlike numeric data, images contain countless attributes. This can lead to difficulties with identifying bias among images [19, 61, 62] or collecting balanced image datasets [63–65].

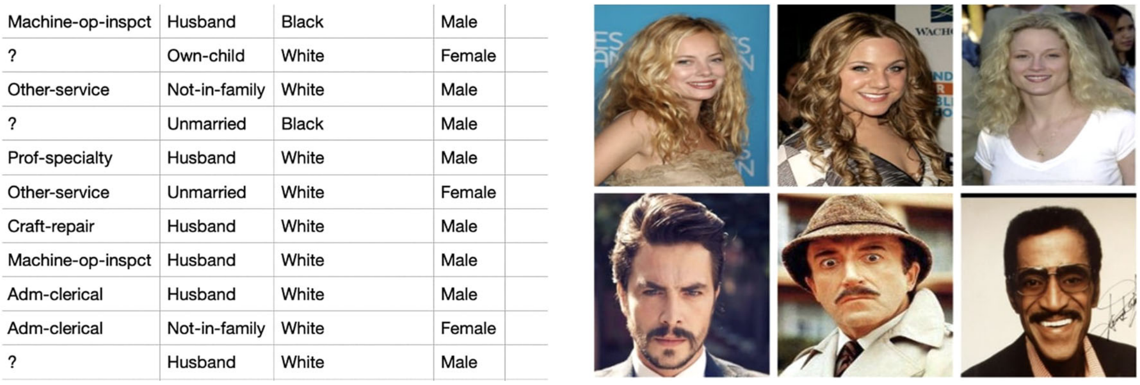
3.2 Problem settings

In this section, we answer the question: What is the research focus for different studies? The section starts with problem settings introduced by training inputs; the problem settings introduced by deep models then follows.

Generally, most image fairness studies focus on fairness-aware classifier training with imbalanced training datasets. Given sensitive features S and a learned model f, a fair learning objective can be expressed as:

Table 2 The datasets most commonly adopted in fairness studies

Datasets	Type	Size	Descriptions
Adult [46]	Numerical	48,842 records	People's salaries with relative information such as age race, gender
COMPAS [47]	Numerical	10,896 records	Criminal defendants' likelihood of reoffending basing on relative attributes
German credit [46]	Numerical	1000 records	Assessments of people as good or bad credit risks described by a set of attributes
Yale B [51]	Image	16,128 images	28 human subjects under 9 poses and 64 illumination conditions
UTKFace [48]	Image	20,000 images	Face images with annotations of age, gender, and ethnicity
CelebA [49]	Image	200k images	Celebrity images, each with 40 attribute annotations
DiF [50]	Image	1M images	Celebrity images, each with 10 attribute annotations



Machine-op-inspct	Husband	Black	Male
?	Own-child	White	Female
Other-service	Not-in-family	White	Male
?	Unmarried	Black	Male
Prof-specialty	Husband	White	Male
Other-service	Unmarried	White	Female
Craft-repair	Husband	White	Male
Machine-op-inspct	Husband	White	Male
Adm-clerical	Husband	White	Male
Adm-clerical	Not-in-family	White	Female
?	Husband	White	Male

Fig. 3 Gender information in the Adult dataset is stored as a binary value [46], while, in the CelebA image dataset, gender information is not stored in a tractable form [49]

$$\min l(Y,\hat{Y}) \quad \text{s.t.} \quad c(f(s_0,s_1))<\theta, \tag{1}$$

where l is the losses, c represents the adopted fairness measurements, and θ represents a small number. The equation illustrates that the trained model should maintain low prediction losses while satisfying some fair measurement constraints with different fairness protection methods. Beyond these general settings, other problem settings have been considered. For example, studies [66–68] consider multiple sensitive attributes: $S=\{S_1,S_2,\ldots\}$, where the sensitive features can have multiple attribute values, such as race. With multiple values, it would be resource-intensive for the fair methods that construct additional sensitive prediction heads. Studies [69, 70] focus on universal fair representations without downstream tasks: $l(I,\hat{I})$, $c(g(s_0,s_1))<\theta$, where g is a feature extraction model, and I and $\hat{I}$ are input images and the generated synthetic images, respectively. This is challenging since it requires the learned features to be independent of the sensitive features. Study [71] consider noisy labels: $(\bar{Y},Y)$, where Y = $\bar{Y}+\epsilon$, $\bar{Y}$ represent the real labels, while Y denotes the labels in the dataset that contain noise. Grari et al. [72] examined continuous sensitive values where $s\in R$. This setting makes the fair methods with sensitive prediction heads invalid since considering prediction heads for continuous values are difficult. Others have concentrated on discovering bias among datasets where the sensitive features S are unknown [61, 62, 73].

Turning to the problem settings introduced by deep models, some studies observed unfair phenomenons in deep models: Choi et al. [69] find that generation models amplify data bias, which leads to unfair image generation. Li et al. [74] consider fair issues in deep clustering methods; Xu et al. [75] observe unfair results in adversarial training, while Chen et al. [76] look at fairness in graph deep models, and Shi et al. [29] summarize image fairness protection methods for deep federated learning models.

One issue that repeatedly crops up in studies on fairness protection concerns model utility once fairness constraints have been enforced. Chang et al. [77], for example, verify that fair models are more vulnerable to adversarial attacks. Mishler and Kennedy investigate the balance between accuracy and fairness [78]. Qian et al. [79] examine fairness methods in the context of deep learning, finding that fair protected models may suffer large fair variance.

Moreover, some studies report that fairness protection issues share connections with issues in other fields. For instance, Zhao et al. [80] cast incremental learning problems to fair protection problems, while Wei et al. [81] explore fairness protection methods for long-tailed data problems. Table 3 summarizes the different problem settings with the corresponding approach.

3.2.1 Discussion

Although images as an input bring certain challenges to fairness protection, they share similar problem settings with numerical studies, such as considering multiple or continuous sensitive attribute values. Similarly, some of the difficulties raised by deep models have also been seen in previous studies, such as fairness for clustering or online learning. However, there are also emerging problems that are being addressed for the first time in the context of image fairness, such as fairness for image generation models, federated models, and graph models.

4 Models of image fairness protections with deep models

While previous studies on fairness protection generally classify deep methods as in-processing methods, we have divided them into four further groups: fair constraint methods, data generation methods, methods that remove sensitive features, and methods that learn independent features. This section introduces each approach in detail and concludes with a comparison of the various characteristics of the methods, including the metrics and datasets used and the sensitive features considered.

4.1 Fair constraints

Fair constraint methods incorporate additional loss constraints and learning objects into the learning procedure in such a way that the learned models should satisfy the corresponding fair metrics. Examples of this method can be found in [75, 77, 82–84] where the learning problems are solved following Eq. 1. This equation can be seen as an optimization issue with constraints. As with other mathematical modeling studies [85–87], the problem can be resolved with Lagrangian methods. Given some sensitive features S and a learning model f with weights w, Eq. 1 can be expressed as:

$$\min_{w} \max_{\lambda \in R_+} l(Y, f(I, w)) + \lambda c(w, S),$$

where λ is the Lagrange multiplier. The maximum lower bound can be replaced considering $\lambda_{\max}$:

$$\max_{\lambda \in R_+} \min_{w} l(Y, f(I, w)) + \lambda c(w, S).$$

The general method of deriving $\lambda_{\max}$ is to update the weights twice during learning, once to minimize the loss w.r.t. w and again to maximize the loss w.r.t. λ at every iteration [88]. Although these methods all rely on constraints, they are designed for different fair metrics and settings. More details on this are presented in Table 5. Notably, while fair constraint methods enforce fairness protection, they do introduce target irrelevant learning objects. To avoid the problem, data operation methods are proposed for better fairness protections. These are discussed next.

4.2 Data operations

Intuitively, there are traditional image processing methods that can enforce fairness. However, studies have shown that general image processing methods are inefficient compared to deep model methods [19]. Thus, some deep models have incorporated techniques like data generation, sampling, and re-weighting to enhance performance.

4.2.1 Data generation

Data generation methods generate synthetic training samples or features with adversarial models to balance the dataset. Hwang et al. [89], for example, generated underrepresented samples with CycleGAN [90], while Joo et al. [65, 91] generated synthetic samples through latent feature manipulations with GAN inversion methods [92]. Alternatively, others have reconstructed the datasets through data argumentation methods [82, 93]. With mixed-up

Table 3 Different problems settings for image fairness protections with deep models

Problems	Specific settings
Most studies focus on general image fairness protection	$\min E[l(Y, \hat{Y})] \quad \text{s.t.} \quad c(f(s_0, s_1)) < \theta$
Problem settings introduced by training data	Settings with sensitive features [66–68, 72] and labels [61, 62, 69, 70, 73]
Problem settings introduced by deep models	Fair for different deep models [29, 69, 74, 75, 75, 76]; Utility studies for fair enforced models [71, 77–79]

images comprised of different subgroups, models tend to learn fair features, like Du et al. [94] for instance, who presented a mix-up scheme to generate neutralized features.

4.2.2 Data sampling

Data sampling methods strike a balance in the training samples during training iterations. In this vein, Roh et al. [30] designed a learning scheme for batch sample selections to balance any prediction gap between subgroups. Another technique is to design deep methods to optimize the sample selection procedure [95, 96]. Shekar et al. [97] enforced fairness through sampling with hard example mining methods.

4.2.3 Data re-weighting

Re-weighting methods introduce parameter weights to balance samples or learned features with the model's design. Zhao et al. [80] reduced biased predictions by attaching one fully connected layer to a classifier, called a weight aligning layer, to re-assign weights across groups. Gong et al. [98] designed different convolution kernels for different attribute values, then re-fused the values to balance feature learning and enforce fairer predictions.

However, while data operations are good for balancing datasets, they do not prevent sensitive features from being learned during training. To overcome this problem, the sensitive features need to be eliminated using a removal method as discussed next.

4.3 Removing sensitive features

Zemel et al. [99] were the first to cast fairness problems as an issue of removing sensitive features from numerical input data. Ever since, researchers have been crafting new ways to remove sensitive features. Some are designed for numerical inputs [38, 100]. Others are designed for image inputs under a range of situations [54, 55, 64, 66, 67, 74, 101–104]. However, despite their subtle differences, all these methods follow the same broad approach. Figure 4 illustrates the general model structure. It is an adversarial framework that learns the target task without the ability to predict sensitive information. Specifically, an encoder first extracts latent information as a proxy for the input. Then, one task prediction head and one sensitive attribute prediction head are attached to predict the learning tasks and the attribute values. To ensure the extracted features contain little to no sensitive information, an inverse gradient is updated for sensitive attribute predictions during the learning. With h representing the extracted features, f and f_s being the task prediction heads and sensitive features prediction heads, respectively, the learning objects can be expressed as:

$$\min_{l} \max_{l_a} E[l(Y,f(h) - l_s(S,f_s(h)],$$

where l and l_s are the training losses of the target and sensitive attribute predictions. The learning adopts an adversarial training setting by minimizing the target classification losses and maximizing the attribute prediction losses.

It is worth noting that, although removing sensitive features enforces fairness, removing features can deteriorate prediction performance. Therefore, the last type of scheme tries to simply disentangle sensitive features without tending to remove them.

4.4 Learning independent features

Methods of learning independent features try to enforce fairness by guaranteeing that the features adopted for task predictions do not contain sensitive information—that is, that the prediction features are independent of the sensitive features. Generally, deep models predict tasks, sensitive attribute values, and reconstruct input data at the same time. A range of similarity measurements between a task's features and its sensitive features have been put forward to enforce independence. The idea is to minimize the similarity between learned features and sensitive features during training. Figure 5 presents the general framework. With the learned features for the

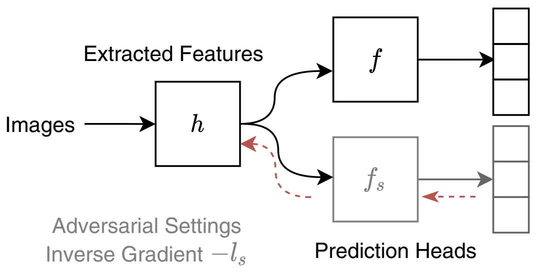


Fig. 4 Adversarial framework with inverse gradient for the methods that remove sensitive features

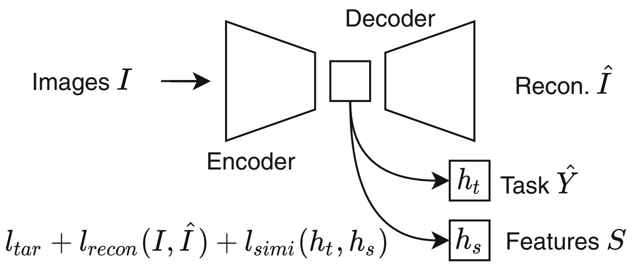


Fig. 5 General model structure for the independent feature learning methods

tasks h_t, the sensitive features h_s, and the reconstructed images $\hat{I}$, the loss functions can be described as:

$$L = \lambda_1 l_{\text{tar}} + \lambda_2 l_{\text{recon}}(I, \hat{I}) + \lambda_3 l_{\text{simi}}(h_t, h_s),$$

where l_{tar} is the target loss, l_{recon} is the reconstruction loss which maintains the utility of generated data, and l_{simi} measures the similarity distance between features.

The studies that have adopted this method in our review include [53, 63, 67, 68, 73, 105–111, 111–117]. The different similarity measurements that have been used include maximum mean discrepancy (MMD) [105] in [105–107], the Kullback–Leibler divergence (KL) [68, 73, 108–111], the Hilbert–Schmidt independence criterion (HSIC) [53, 112, 113], matrix correlations [114], cosine similarity [63, 115, 116], and L1 and Euclidean distance [67, 117]. Additionally, Gitiaux and Rangwala [111] designed a binary representation method to better learn independent features. Table 4 provides a detailed summary of the main methods referred to in this section.

4.4.1 Other studies

Apart from deep model designs, there are other technologies for enforcing image fairness protection. For example, Wang and Deng [118] learn adaptive classification margins for different subgroups with deep reinforcement learning methods (DQN [119]), while Kim et al. [120] adopt boosting methods to promote fairness.

4.4.2 Discussion

Although above methods are widely adopted for fairness-aware training, there are limitations for each direction. Using fair constraints usually results in an accuracy drop, often referred to as the utility issue, which has been reported in several studies such as [35, 121, 123, 124]. Moreover, deep model optimizations are often non-convex in nature and challenging to solve in general, which may lead to difficult or unstable training [26, 125].

With the data generation techniques, generating under-represented images or counterfactual samples for model training is difficult as disentangling features and interpreting high dimensional inputs is always challenging [92]. Additionally, Vinyals [126, 127] show that, even generating highly photorealistic images, e.g., with model BigGAN [128], training a classifier on synthetic images is never as good as training with real ones.

Removing sensitive features and learning independent features both involve representation learning, which raises similar concerns over utility [19, 110, 117]. It also creates difficulties when attempting to learn invariant or independent features [102]. In their survey, Caton and Haas [26] discuss the limitations of fair representation learning studies in detail.

Despite all the different directions pursued, each aims to break spurious correlations between learning targets and sensitive features. Fair constraint methods encourage models to learn similar predictions across subgroups of sensitive features, which leaves the predictions and sensitive features irrelevant. Data operation methods balance the datasets and reweight the feature maps to impede the model's short-cut learning. Sensitive feature removal prevents the models from extracting sensitive features through deep adversarial designs. And learning independent features encourages the model predictions to be independent of the sensitive features through various similarity measurements. Nevertheless, all four methods aim to generate fair predictions by not learning spurious correlations.

4.4.3 Fairness characteristics

Having discussed the methods themselves, it is now important to cover the measurements, datasets, and attributes considered for a more detailed comparison between the methods. Table 5 shows a matrix. The table illustrates that most studies concentrate on a single fairness metric, while strategies that involve multiple metrics have drawn less attention. Additionally, there are fewer studies on individual or counterfactual fairness since the metrics required are stricter than for others. Another issue that has been raised is that the frameworks are not particularly generalized. A method designed for one fair metric may not be applicable to others [79]. Hence, general frameworks for fairness are still necessary. As a last observation, most of the studies concern data with a single sensitive feature, such as gender or race. Problems involving multiple sensitive features or features with non-binary or continuous values are far less studied.

In real applications, given limited medical data available for training, Frid-Adar et al. [129] opted to generate their own data and achieved outstanding fairness protection results. After comparing different fairness protection methods, Qian et al. [79] found that fair constraint methods were suitable for fairness protection under various fair metrics since the constraints are explicitly attached to the loss functions. Wang et al. [19] find that, in general, representation methods such as removing sensitive features and learning independent features can enforce fairness protection. However, learning fair representations is difficult.

Overall, despite the different methods, the current studies have only eased the unfair prediction issue, not

Table 4 A comparison of deep learning image fairness models

Names	Models	Outputs	Targets	Remv.	Recon.	Simil.	Fair const
ALFR [54]	Encoder	Pred.	w	w	w/o	w/o	w/o
Beutel [55]	Encoder	Pred.	w	w	w/o	w/o	w/o
FFVAE [109]	VAE	Data	w	w/o	w	w	w/o
AFR [103]	GAN	Data	w	w/o	w	w/o	w
Zhang [101]	Encoder	Pred.	w	w	w/o	w/o	w/o
FAD [121]	Encoder	Pred.	w	w	w/o	w/o	w/o
FairGAN [52]	GAN	Data	w	w	w/o	w/o	w/o
Quadrianto [53]	GAN	Data	w	w/o	w	w	w/o
Louizos [107]	VAE	Data	w	w/o	w	w	w/o
Moyer [110]	VAE	Data	w	w/o	w	w	w/o
FCNN [84]	Encoder	Pred.	w	w/o	w/o	w/o	w
FRL [75]	Encoder	Pred.	w	w/o	w/o	w/o	w
Chang [77]	Encoder	Pred.	w	w/o	w/o	w/o	w
FairMixup [82]	Encoder	Pred.	w	w/o	w/o	w/o	w
ApxFair [83]	Encoder	Pred.	w	w/o	w/o	w	w/o
Joo [91]	GAN	Data	w/o	w/o	w	w/o	w/o
Wang [19]	Encoder	Pred.	w	w	w/o	w/o	w/o
Wang [64]	Encoder	Pred.	w	w/o	w/o	w/o	w
Ramaswamy [65]	GAN	Data	w/o	w/o	w	w/o	w/o
FairFaceGAN [66]	GAN	Data	w	w	w	w	w/o
Hwang [67]	Encoder	Pred.	w	w	w/o	w	w/o
Xie [102]	Encoder	Pred.	w	w	w/o	w/o	w/o
DB-VAE [73]	VAE	Data	w	w/o	w	w	w/o
FD-VAE [108]	VAE	Data	w	w	w/o	w	w/o
Sarhan [68]	Encoder	Pred.	w	w/o	w	w	w/o
MFD [106]	Encoder	Pred.	w	w/o	w/o	w	w/o
Tartaglione [114]	Encoder	Pred.	w	w/o	w/o	w	w/o
Xu [117]	Encoder	Pred.	w	w/o	w/o	w	w/o
Boedi [22]	Encoder	Pred.	w	w/o	w/o	w/o	w
CSAD [115]	Encoder	Pred.	w	w/o	w/o	w	w/o
Mo [93]	Encoder	Pred.	w	w/o	w/o	w	w/o
BiasCon [116]	Encoder	Pred.	w	w/o	w/o	w	w/o
Boedi [122]	Encoder	Pred.	w	w/o	w/o	w	w/o

Pred. = prediction results. Data = features or synthetic data. The objective functions include: target loss, sensitive removing loss, reconstruction loss, feature similarity loss, and fair constraint loss

solved it. Fixing all the problems with fairness protection still holds many challenges. This is the subject of the next section. However, it is notable that although the fair results may just be improved 2% in the studies, the studies are working towards solving the issue and better understanding the dynamic learning procedures of deep models.

5 Challenges to ensuring greater fairness

This section discusses the challenges underlying the above methods. From our review, we settled on three main questions that need to be answered before fairness protection can become a largely resolved issue: (1) How can we learn fairer representations? (2) How can we maintain utility in the face of data modifications (data operation methods) or additional training objects (fair constraints)? And (3) Since most studies concern public datasets, how can we reveal imbalance bias in real data and real-world situations? Each of these questions is discussed in more detail next.

5.1 How to get fairer representations?

The critical issue for the methods that involve removing sensitive features or learning independent features with

Table 5 A comparison of the main fair characteristics

Names	Dataset	Features	Metrics	Methods	Descriptions
ALFR [54]	Adult	Gender	DP	SFR	With attributes classifiers to remove sensitive features
Beutel [55]	Adult	Gender	DP	SFR	Proposed representation for fairness
FFVAE [109]	CelebA	Chubby, Egyglasses, Gender	DP	IFL	Fairness with multiple-sensitive features
AFR [103]	Adult	Gender	DP	SFR	Removed features with attributes classifications and reconstruction loss
Zhang [101]	UCI Adult	Gender	DP EOD EOP	SFR	Removed sensitive features without intermediate representations
FAD [121]	Adult	Gender	DP DM	SFR	Further integrated models in [103] into one GAN based model
FairGAN [52]	UCI Adult	Gender	DP DM	SFR	First introduced GAN based FairGAN to enforce fairness and data generation
Quadrianto [53]	CelebA	Gender	EOP	SFR	Proposed GAN based models to enforce fairness.
Louizos [107]	YaleB	Illumination	DP EOP IF	IFL	MMD to measured distance of features
FCNN [84]	Adult	gender	DP EOD	FC	With Lagrangian to solve fair constraints
Chang [77]	Adult	Gender	ACC	FC	Attached fairness constraints to attack objectives
FairMixup [82]	CelebA	Gender	DP EOD	FC	With mixup images and proposed new constraints for fairness
ApxFair [83]	Adult	Gender	DP EOD	FC	Developed iterative algorithms to solve fair constraints.
Joo [91]	CelebA	Gender	ACC	DO	Generated counterfactual samples for invariant features learning to enforce fairness
Wang [19]	CelebA	Gender	BA	DO	Trained with ALL domains for features independent learning
Ramaswamy [65]	CelebA	Gender	BA	DO	Generated counterfactuals with latent vectors to enforce fairness
Hwang [89]	CelebA	Gender	ACC	DO	Based on cycleGAN to generate fair images.
Xie [102]	Yale B	Illumination	ACC	SFR	Removed sensitive features with attributes classifiers
Wang [64]	Coco [130]	Gender	BA	SFR	Required similar outputs with environments / backgrounds
FairFaceGAN [66]	CelebA	Gender	EOD EOP	SFR	Removed features with multiple sensitive attributes classifiers
Hwang [67]	CelebA	Gender	EOP	SFR	With Triplet and Group loss to measure feature distance
DB-VAE [73]	CelebA	Race Gender	ACC	IFL	Learned the underlying latent variables through minimizing KL distance
FD-VAE [108]	CelebA UTK	Gender Race	EOD EOP	IFL	Adopted VAE distribution to disentangle and decorrelated features
Sarhan [68]	YaleB	Illumination	ACC	IFL	Designed disentanglement loss with mean orthogonal to learn fair representations
MFD [106]	CelebA UTK	Gender Race	EOD	IFL	Used MMD to learn independent features
Tartaglione [114]	CelebA	Gender	ACC	IFL	Measured feature distance with matrix correlations
Xu [63]	BUPT [131]	Race	ACC	IFL	Measured distance with Cosine Similarity Matrix
Boedi [117]	CelebA	Gender	EOP	IFL	Adopted L1 Norm to measure feature distances
CSAD [115]	CelebA	Gender	ACC	IFL	Learned disentangled features with cosine similarity constraints
Mo [93]	Coco	Scene	ACC	DG	With ContraCAM and data augmentations to debias scene correlations
BiasCon [116]	CelebA UTK	gender race	ACC	IFL	With contrastive learning based losses to alleviate bias predictions

Table 5 (continued)

Names	Dataset	Features	Metrics	Methods	Descriptions
Boedi [122]	CelebA	gender race	EOP DP	IFL	Diminished feature distance across groups to enforce fairness

DP demographic parity, *SP* statistical parity, *EOD* equalized odds, *EOP* equal opportunity, *DM* disparate mistreatment [132], *IF* individual fairness, *BA* bias amplification [133]. Methods: *DO* data operations, *FC* fair constraints, *SFR* sensitive feature removal, *IFL* independent features learning

similarity constraints is to learn fair representations. Although previous studies have proposed solutions like gradient reversing methods or learning independent features, solutions from other fields experiencing similar problems may also be helpful. We reviewed some of the literature on image causality inference, domain adaptation (invariant representation learning), and transfer learning and found several methods worthy of discussion that could be adopted to improve fair representation learning.

5.1.1 Image causality inference

Image causality inference is an emerging topic. The idea behind this notion is to empower learning models with the ability to deal with causal effects; they can either remove the spurious bias [134], disentangle the desired model effects [135], or modularize reusable features that generalize well [136]. The main approach involves interventions with inputs or features. At the same time, similar predictions are required for original and modified inputs. As a result, causal chains are broken down for skew attributes and targets. Given the original and counterfactual samples x and $\hat{x}$ with a learning target of Y, the consistency rule is formulated as:

$$P(Y|x) = P(Y|\hat{x}),$$

where P are the prediction probabilities. Most studies construct a contrastive loss to ensure similar predictions. For the training pair x and $\hat{x}$ with a learning model f, the contrastive loss can be expressed as:

$$L_{\text{con}} = -\log E\left[\frac{\exp(f(x)^T f(\hat{x})/\tau)}{\sum_i \exp(f(x_i)^T f(x_j)/\tau)}\right],$$

where x_i and x_j are samples from different classes and τ is the scalar temperature hyper-parameter [137].

The connection between causality and fairness has been shown in several studies. Kusner et al. introduced counterfactual fairness, which requires that samples and their counterfactual counterparts should share similar predictions [33]. Ramaswany et al. and Yurochkin et al. generated synthetic images by manipulating latent vectors as counterfactual samples [36, 65]. Sarhan et al. treated sensitive features S and learning targets Y as causally independent features through learning orthogonality values of mean vectors for target and sensitive features distributions [68].

5.1.2 Domain adaptation

Domain adaptation, also known as invariant representation learning, refers to methods that attempt to learn invariant features across domains for the purposes of model generalization improvement. The invariant learning serves as a proxy for causality inference [138], it refers to finding features that are domain-invariant, i.e, that reliably predict the true class regardless of the domain environment [139]. As domains can be viewed as sensitive group features in fairness protection, fairness problems can be cast as an issue of learning invariant features. Invariant risk minimization [140] is one of several recently successful approaches in the field that minimizes predicting distances across domains. Given the domains $\forall e_1, e_2 \in E$ through and the learning features $\forall h \in H$, the aim of these studies can be expressed as:

$$E[y \mid g(x) = h, e_1] = E[y \mid g(x) = h, e_2],$$

where g are the feature extraction models.

Requiring invariant features and consistency predictions across domains can be interpreted as a group fairness metric, such as demographic fairness. Similar interpretations have been discussed in [141] which enumerates several group fairness criteria and draws analogies to domain generalization methods. For fairness protection, Adragna et al. empirically demonstrate that domain adaptation methods can be used to enforce fairness protection through learning models that are invariant to the features containing sensitive attributes [142]. They adopted invariant risk minimization to encourage models to learn invariant predictors for different sensitive subgroups. Although they considered only textual inputs for comment toxicity classification tasks, in principle, the proposed methods could be applied to tasks with images.

5.1.3 Transfer learning

Some image fairness methods enforce fairness by requiring similar feature distributions across subgroups

[68, 73, 108, 109]. Although the problem may not be the same as transfer learning, which requires transferring the knowledge across domains, they share similar methods such as MMD in [106, 107], KL in [73], or HSIC constraint in [53]. We think the merging methods in transfer learning can be further adapted to encourage better fairness protection.

5.1.4 Other fields

There are also other fields that closely relate to fairness protection, such as: out-of-distribution detection (OOD) [143], which distinguishes minorities based on feature differences; GAN inversion [65], which inverts images into disentangled latent space features; contrastive learning [144], which requires consistent predictions for training pairs; incremental learning [80], which manipulates the learning features for different classes. Such methods might also be helpful for encouraging fair representation learning and fairness protection.

5.2 How to maintain prediction performance after enforcing fairness?

Fairness methods may modify training samples or introduce target-irrelevant constraints. Naturally, this raises concerns about whether the applying methods will cause the model's performance to deteriorate. Most studies on enforcing fairness in deep models witness accuracy drops after fairness protection has been applied [121, 123]. A similar phenomenon has also been observed in traditional machine learning studies [35, 124], so this comes as no surprise. Further, Chang et al. [77] observed that fair models are more vulnerable to adversarial attacks than their original counterparts. Van et al. [145] discuss adversarial defenses for fair models. Qian et al. [79] illustrate that learned models tend to have larger fair variance after fairness enforcement. In other words, they have utility problems. Few studies have focused on this problem, so this is a future challenge still to be met.

5.2.1 Reasons

Currently, no closed studies are available to illustrate the reason for utility issues. The issue may be caused by removing sensitive features, which may actually remove the features related to the targets [40] for methods that remove sensitive features or it could be due to any target irrelevant constraints, such as the additional losses introduced in fair constraint methods.

5.2.2 Methods

To maintain utility, one promising direction is to maintain the similarities between original and learned data. Calmon et al. [38] introduced a utility preservation constraint to guarantee that the distributions between the original data and the latent space features remained statistically close. Specifically, they adopted KL-divergence to measure the distances between two distributions. Zhang et al. measured the same distance but in Euclidean terms [100], while Xu et al. use dimension-wise probability to check whether the modified data maintains a similar distribution [52]. Beyond distribution similarities, Quadrianto et al. adopt image reconstruction losses to ensure similar semantic meanings for generated images [53].

Previous studies concentrate on maintaining similarity to guarantee the utility of fairness protection methods. However, proposing a general and practical metric to measure utility is still challenging, especially for image modifications. Additionally, the introduced methods have only been applied to datasets with limited samples under specific settings. Thus, the methods' performance could be positively related to factors such as the size of data, the sparsity of data, the data that is modified, or the specific machine learning algorithm used. Further study on the utility impact issue is still necessary.

5.3 How to find imbalanced biases among data

Since most studies on fairness protection rely on public datasets or synthetically generated data with given sensitive feature annotations, more work is needed to determine how we can effectively discover bias in real-world data. Hu et al. introduced the one-human-in-the-loop method to find bias [62]. They designed questions to ask people whether images contained sensitive information, revealing sensitive attributes based on the statistics. Amini et al. discovered under-represented data based on the latent representations [73]. They distinguished feature representations between majority and minority data, which is challenging with image data. Li et al. [61] combined the methods and tried to find biased features through GAN and human-in-the-loop methods. They generated synthetic images with GAN methods based on the latent vectors and asked people to interpret the semantic meaning of images.

As few studies focus on finding bias, considering the high cost of mass data annotations, more methods to discover bias in datasets are necessary.

6 From fairness to privacy

6.1 Fairness protection and privacy protection share relationships

Dwork et al. introduced the notion of individual fairness, which requires treating similar people similarly [31]. The idea can be seen as a generalization of differential privacy [146]. As described in [99], differential privacy involves a constraint for the rows of a data matrix, while fairness involves a constraint for the columns. They share tide relations. Inspired by the similarity between the two, Dwork et al. [31] proposed a fairness protection method based on differential privacy that imposes a Lipschitz constraint on fair metrics. Likewise, fairness protection methods have also been considered for privacy protection issues [147].

6.2 With machine learning

Ekstrand et al. first raised the question of whether statistical metrics of predictive results, such as equalized odds, were compatible with privacy [148]. They showed that the constraint of differential privacy might lead to fairness under certain conditions. Later, Jagielski et al. proposed two algorithms that can satisfy both differential privacy and equalized odds. Similarly, Xu et al. also achieved differential privacy and demographic parity at the same time [149]. Other studies include [150] and [151] with K-anonymity, and [152] and [153] with data mining.

6.3 With deep learning

New challenges have emerged from the studies that focus on fairness protection in deep learning. Xu et al. and Bagdasaryan et al., for instance, find that privacy protection with stochastic gradient descent may lead to unfair results [123, 154]. This shows that achieving fairness protection and differential privacy at the same time is quite necessary. As for fair representation, Grgic-Hlaca et al. [24] regard the sensitive features of fairness as private information and so proposed methods that fit both fairness and privacy. Edwards et al. [54] use learned representations to hide private information in the image. They argue that, in this way, image privacy and fairness can be achieved at the same time.

6.4 Future directions and conclusions

There are several outstanding challenges in the image fairness literature that have yet to be addressed. In Sect. 3, we indicated that future trends in this field may fall into: (1) exploring more and different settings, such as multiple [66] and continuous [72] sensitive features; and (2) examining some newly emerging deep learning applications, such as deep clustering [74], adversarial training [75], and attacks [77].

Given the models and methods presented in Sects. 4 and 5, more studies from related study fields could be taken on board to ensure fairer representations. Some solutions already seem very promising, such as causality inference [65], domain generalization or invariant representation learning [142], and transfer learning [103].

In terms of the model utility concerns presented in Sect. 5, a systematic understanding of how data and feature modifications contribute to predictions is still lacking. We believe the causal-based [155, 156] methods or those based on interpretation [157] could be possible solutions. Moreover, as little attention has been paid to discovering sensitive features, further research is required to discover and represent them in real-world datasets [24].

Additionally, fairness protection and privacy protection have a great many overlaps. Yet, few methods have been proposed to achieve both [123, 154, 158]. Further research should be undertaken to explore methods that can bestow both types of protection, especially for settings with high dimensional inputs.

In this paper, we summarized deep learning based image fairness protection studies in three respects: problems, models, and challenges. Since image inputs are different from numerical inputs, we started by highlighting the differences and summarizing the research trends by presenting different problem settings. We then introduced four approaches to fairness with deep model methods and their characteristics. Additionally, we discussed the three main challenges leading to better fairness protection results. Last, we discussed the closeness between fairness and privacy as issues.

Although our focus remained solely fixed on image fairness studies in the realm of deep learning, we did introduce some problems and challenges that can extend to fairness with numerical inputs and other high dimensional data tasks, such as natural language processing, speech processing, or video processing. Problems with data bias are also common in the image processing area. The discussed studies shared the same problems with other fields such as domain adaptation, transfer learning, or long-tail issues, as all aim to break spurious correlations and learn invariant features. We expect that the methods might stimulate cross-pollination among these fields.

Our survey concludes with some comparisons between fairness and privacy preservation in terms of problems and questions. Fairness protection methods can be aligned with accuracy and privacy preservation. This leaves room for further work to summarize fairness protection studies from a range of additional perspectives.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest The authors (from University of Technology Sydney and the City University of Macau) declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Guion R (2008) Employment tests and discriminatory hiring. Ind Relat A J Econ Soc 5:20–37. https://doi.org/10.1111/j.1468-232X.1966.tb00449.x
2. Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia PK, Mehta S, Mojsilovic A, Nagar S, Ramamurthy KN, Richards JT, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y (2019) Think your artificial intelligence software is fair? Think again. IEEE Softw 36:76–80
3. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2018) Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions. arXiv preprint arXiv:181107867
4. Malekipirbazari M, Aksakalli V (2015) Risk assessment in social lending via random forests. Expert Syst Appl 42(10):4621–4631
5. Hoffman M, Kahn LB, Li D (2018) Discretion in hiring. Q J Econ 133(2):765–800
6. Perlich C, Dalessandro B, Raeder T, Stitelman O, Provost F (2014) Machine learning for targeted display advertising: transfer learning in action. Mach Learn 95(1):103–127
7. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8–17
8. Oneto L, Donini M, Pontil M, Maurer A (2020) Learning fair and transferable representations with theoretical guarantees. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA). IEEE, pp 30–39
9. Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Trans Inf Syst (TOIS) 14(3):330–347
10. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. arXiv preprint arXiv:190809635
11. Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social data: biases, methodological pitfalls, and ethical boundaries. Front Big Data 2:13
12. Suresh H, Guttag JV (2019) A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:190110002
13. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE/CVF conference on computer vision and pattern recognition (CVPR)
14. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition, pp 3354–3361
15. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency, PMLR, pp 77–91
16. Brandao M (2019) Age and gender bias in pedestrian detection algorithms. arXiv preprint arXiv:190610490
17. Benthall S, Haynes BD (2019) Racial categories in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency
18. Hanna A, Denton EL, Smart A, Smith-Loud J (2020) Towards a critical race methodology in algorithmic fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency
19. Wang Z, Qinami K, Karakozis IC, Genova K, Nair P, Hata K, Russakovsky O (2020) Towards fairness in visual recognition: effective strategies for bias mitigation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8919–8928
20. Wang M, Deng W (2020) Mitigating bias in face recognition using skewness-aware reinforcement learning. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR) pp 9319–9328
21. Mo S, Kang H, Sohn K, Li CL, Shin J (2021) Object-aware contrastive learning for debiased scene representation. Adv Neural Inf Process Syst 34
22. Xu X, Huang Y, Shen P, Li S, Li J, Huang F, Li Y, Cui Z (2021) Consistent instance false positive improves fairness in face recognition. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 578–586
23. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency, pp 59–68
24. Grgic-Hlaca N, Redmiles EM, Gummadi KP, Weller A (2018) Human perceptions of fairness in algorithmic decision making: a case study of criminal risk prediction. In: Proceedings of the 2018 world wide web conference, international world wide web conferences steering committee, pp 903–912
25. Quy TL, Roy A, Iosifidis V, Ntoutsi E (2021) A survey on datasets for fairness-aware machine learning. arXiv:2110.00530
26. Caton S, Haas C (2020) Fairness in machine learning: a survey. arXiv preprint arXiv:201004053
27. Malik N, Singh PV (2019) Deep learning in computer vision: methods, interpretation, causation and fairness. Interpretation, causation and fairness (May 28, 2019)
28. Du M, Yang F, Zou N, Hu X (2021) Fairness in deep learning: a computational perspective. IEEE Intell Syst 36:25–34
29. Shi Y, Yu H, Leung C (2021) A survey of fairness-aware federated learning. arXiv:2111.01872
30. Roh Y, Lee K, Whang SE, Suh C (2021) Fairbatch: batch selection for model fairness. arXiv:2012.01696

31. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp 214–226
32. Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. In: Advances in neural information processing systems
33. Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Advances in neural information processing systems, pp 4066–4076
34. Zhang J, Bareinboim E (2018) Fairness in decision-making-the causal explanation formula. In: Thirty-second AAAI conference on artificial intelligence
35. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp 3315–3323
36. Yurochkin M, Sun Y (2021) Sensei: sensitive set invariance for enforcing individual fairness. In: International conference on learning representations
37. Islam MT, Fariha A, Meliou A (2021) Through the data management lens: experimental analysis and evaluation of fair classification. arXiv preprint arXiv:210107361
38. Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR (2017) Optimized pre-processing for discrimination prevention. In: Proceedings of the 31st international conference on neural information processing Systems, pp 3995–4004
39. Iosifidis V, Ntoutsi E (2018) Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D Clough Robert Jäschke, pp 24–29
40. Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33
41. Hazan E (2019) Introduction to online convex optimization. arXiv preprint arXiv:190905207
42. Redmond M, Baveja A (2002) A data-driven software tool for enabling cooperative information sharing among police departments. Eur J Oper Res 141(3):660–678
43. Chen R, Lucier B, Singer Y, Syrgkanis V (2017) Robust optimization for non-convex objectives. arXiv preprint arXiv:170701047
44. Freund Y, Schapire RE (1999) Adaptive game playing using multiplicative weights. Games Econom Behav 29(1–2):79–103
45. Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A (2016) The case for process fairness in learning: feature selection for fair decision making. In: NIPS symposium on machine learning and the law, vol 1, p 2
46. Asuncion A, Newman D (2007) UCI machine learning repository
47. Larson J, Mattu S, Kirchner L, Angwin J (2016) Compas analysis. GitHub, available at: https://github.com/propublica/compas-analysis
48. Geralds J (2017) Utkface large scale face dataset. github.com
49. Liu Z, Luo P, Wang X, Tang X (2018) Large-scale celebfaces attributes (celeba) dataset. Retrieved August 15, 2018
50. Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. arXiv preprint arXiv:190110436
51. Lee KC, Ho J, Kriegman D (2005) The extended yale face database b. Online] http://vision.ucsd.edu/leekc/ExtYaleDatabase/ExtYaleB.html
52. Xu D, Yuan S, Zhang L, Wu X (2018) Fairgan: fairness-aware generative adversarial networks. In: 2018 IEEE international conference on big data (Big Data). IEEE, pp 570–575
53. Quadrianto N, Sharmanska V, Thomas O (2019) Discovering fair representations in the data domain. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8227–8236
54. Edwards H, Storkey A (2015) Censoring representations with an adversary. arXiv preprint arXiv:151105897
55. Beutel A, Chen J, Zhao Z, Chi EH (2017) Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint arXiv:170700075
56. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114
57. Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:14014082
58. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
59. d'Alessandro B, O'Neil C, LaGatta T (2017) Conscientious classification: a data scientist's guide to discrimination-aware classification. Big data 5(2):120–134
60. Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, et al. (2018) Ai fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:181001943
61. Li Z, Xu C (2021) Discover the unknown biased attribute of an image classifier. arXiv preprint arXiv:210414556
62. Hu X, Wang H, Vegesana A, Dube S, Yu K, Kao G, Chen SH, Lu YH, Thiruvathukal GK, Yin M (2020) Crowdsourcing detection of sampling biases in image datasets. Proc Web Conf 2020:2955–2961
63. Xu X, Huang Y, Shen P, Li S, Li J, Huang F, Li Y, Cui Z (2021) Consistent instance false positive improves fairness in face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 578–586
64. Wang T, Zhao J, Yatskar M, Chang KW, Ordonez V (2019) Balanced datasets are not enough: estimating and mitigating gender bias in deep image representations. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5310–5319
65. Ramaswamy VV, Kim SS, Russakovsky O (2021) Fair attribute classification through latent space de-biasing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9301–9310
66. Hwang S, Park S, Kim D, Do M, Byun H (2020) Fairfacegan: fairness-aware facial image-to-image translation. In: 31st British machine vision conference 2020, BMVC
67. Hwang S, Park S, Lee P, Jeon S, Kim D, Byun H (2021) Exploiting transferable knowledge for fairness-aware image classification. In: Ishikawa H, Liu CL, Pajdla T, Shi J (eds) Computer vision—ACCV 2020, vol 12625. Springer, Cham, pp 19–35. https://doi.org/10.1007/978-3-030-69538-5_2
68. Sarhan MH, Navab N, Eslami A, Albarqouni S (2020) Fairness by learning orthogonal disentangled representations. In: European conference on computer vision. Springer, pp 746–761
69. Grover A, Choi K, Shu R, Ermon S (2020) Fair generative modeling via weak supervision. In: ICML
70. Kairouz P, Liao J, Huang C, Sankar L (2019) Censored and fair universal representations using generative adversarial models. arXiv preprint arXiv:191000411
71. Celis LE, Huang L, Keswani V, Vishnoi NK (2021) Fair classification with noisy protected attributes: a framework with provable guarantees. In: International conference on machine learning. PMLR, pp 1349–1361
72. Grari V, Ruf B, Lamprier S, Detyniecki M (2019) Fairness-aware neural r\'eyni minimization for continuous features. arXiv preprint arXiv:191104929
73. Amini A, Soleimany AP, Schwarting W, Bhatia SN, Rus D (2019) Uncovering and mitigating algorithmic bias through

learned latent structure. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 289–295

74. Li P, Zhao H, Liu H (2020) Deep fair clustering for visual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition p 10

75. Xu H, Liu X, Li Y, Jain A, Tang J (2021) To be robust or to be fair: towards fairness in adversarial training. In: International conference on machine learning, PMLR, pp 11492–11501

76. Chen D, Lin Y, Zhao G, Ren X, Li P, Zhou J, Sun X (2021) Topology-imbalance learning for semi-supervised node classification. Adv Neural Inf Process Syst 34

77. Chang H, Nguyen TD, Murakonda SK, Kazemi E, Shokri R (2020) On adversarial bias and the robustness of fair machine learning. arXiv preprint arXiv:200608669

78. Mishler A, Kennedy EH (2021) Fade: Fair double ensemble learning for observable and counterfactual outcomes. arXiv:2109.00173

79. Qian S, Pham H, Lutellier T, Hu Z, Kim J, Tan L, Yu Y, Chen J, Shah S (2021) Are my deep learning systems fair? An empirical study of fixed-seed training. Adv Neural Inf Process Syst 34

80. Zhao B, Xiao X, Gan G, Zhang B, Xia ST (2020) Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13208–13217

81. Wei C, Sohn K, Mellina C, Yuille A, Yang F (2021) Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10857–10866

82. Chuang CY, Mroueh Y (2021) Fair mixup: fairness via interpolation. In: International conference on learning representations

83. Mandal D, Deng S, Jana S, Wing J, Hsu DJ (2020) Ensuring fairness beyond the training data. Adv Neural Inf Process Syst 33:18445–18456

84. Manisha P, Gujar S (2020) FNNC: Achieving fairness through neural networks. In: IJCAI

85. Al-Smadi M, Arqub OA, Hadid SB (2020) An attractive analytical technique for coupled system of fractional partial differential equations in shallow water waves with conformable derivative. Commun Theor Phys 72:085001

86. Alabedalhadi M, Al-Smadi M, Al-Omari S, Baleanu D, Momani S (2020) Structure of optical soliton solution for nonlinear resonant space–time schrödinger equation in conformable sense with full nonlinearity term. Phys Scr 95:105215

87. Al-Smadi M, Arqub OA, Momani S (2020) Numerical computations of coupled fractional resonant schrödinger equations arising in quantum mechanics under conformable fractional derivative sense. Phys Scr 95:075218

88. Eban E, Schain M, Mackey A, Gordon A, Rifkin R, Elidan G (2017) Scalable learning of non-decomposable objectives. In: Artificial intelligence and statistics. PMLR, pp 832–840

89. Hwang S, Byun H (2020) Unsupervised image-to-image translation via fair representation of gender bias. In: ICASSP 2020—2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Barcelona, Spain, pp 1953–1957. https://doi.org/10.1109/ICASSP40776.2020.9054129

90. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

91. Joo J, Kärkkäinen K (2020) Gender slopes: counterfactual fairness for computer vision models by attribute manipulation. In: Proceedings of the 2nd international workshop on fairness, accountability, transparency and ethics in multimedia, pp 1–5

92. Xia W, Zhang Y, Yang Y, Xue JH, Zhou B, Yang MH (2021) Gan inversion: a survey. arXiv preprint arXiv:210105278

93. Mo S, Kang H, Sohn K, Li CL, Shin J (2021) Object-aware contrastive learning for debiased scene representation. arXiv:2108.00049

94. Du M, Mukherjee S, Wang G, Tang R, Awadallah A, Hu X (2021) Fairness via representation neutralization. Adv Neural Inf Process Syst 34

95. Khalili MM, Zhang X, Abroshan M (2021) Fair sequential selection using supervised learning models. arXiv:2110.13986

96. Bendekgey H, Sudderth E (2021) Scalable and stable surrogates for flexible classifiers with fairness constraints. Adv Neural Inf Process Syst 34

97. Shekhar S, Ghavamzadeh M, Javidi T (2021) Adaptive sampling for minimax fair classification. arXiv:2103.00755

98. Gong S, Liu X, Jain AK (2021) Mitigating face recognition bias via group adaptive classifier. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3414–3424

99. Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: International conference on machine learning, pp 325–333

100. Zhang L, Wu Y, Wu X (2017) Achieving non-discrimination in data release. In: Proceedings of the 23rd ACM SIGKDD International conference on knowledge discovery and data mining. ACM, pp 1335–1344

101. Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society. ACM, pp 335–340

102. Qizhe H, Zihang D, Yulun D, Hovy E, Neubig G (2017) Controllable invariance through adversarial feature learning. In: Isabelle G, Ulrike von L, Samy B, Hanna M. W, Rob F, Vishwanathan S. V. N, Roman G (eds) Advances in neural information processing systems 30. NIPS, USA, pp 585–596

103. Madras D, Creager E, Pitassi T, Zemel R (2018) Learning adversarially fair and transferable representations. In: International conference on machine learning. PMLR, pp 3384–3393

104. Robinson JP, Qin C, Henon Y, Timoner S, Fu YR (2021) Balancing biases and preserving privacy on balanced faces in the wild. arXiv:2103.09118

105. Smola AJ, Gretton A, Borgwardt K (2006) Maximum mean discrepancy. In: 13th international conference, ICONIP 2006, Hong Kong, China, October 3–6, 2006: proceedings

106. Jung S, Lee D, Park T, Moon T (2021) Fair feature distillation for visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12115–12124

107. Louizos C, Swersky K, Li Y, Welling M, Zemel R (2015) The variational fair autoencoder. arXiv preprint arXiv:151100830

108. Park S, Kim D, Hwang S, Byun H (2020) Readme: Representation learning by fairness-aware disentangling method. arXiv preprint arXiv:200703775

109. Creager E, Madras D, Jacobsen JH, Weis MA, Swersky K, Pitassi T, Zemel R (2019) Flexibly fair representation learning by disentanglement. arXiv preprint arXiv:190602589

110. Daniel M, Shuyang G, Rob B, Aram G, Greg Ver S (2018) Invariant representations without adversarial training. In: Samy B, Hanna M. W, Hugo L, Kristen G, Roman G (eds) Advances in neural information processing systems 31. NIPS, Canada, pp 9102–9111

111. Gitiaux X, Rangwala H (2021) Fair representations by compression. In: AAAI

112. Gretton A, Bousquet O, Smola A, Schölkopf B (2005) Measuring statistical dependence with Hilbert–Schmidt norms. In:

International conference on algorithmic learning theory. Springer, pp 63–77

113. Bahng H, Chun S, Yun S, Choo J, Oh SJ (2020) Learning debiased representations with biased representations. In: ICML
114. Tartaglione E, Barbano CA, Grangetto M (2021) End: entangling and disentangling deep representations for bias correction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13508–13517
115. Zhu W, Zheng H, Liao H, Li W, Luo J (2021) Learning bias-invariant representation by cross-sample mutual information minimization. arXiv:2108.05449
116. Hong Y, Yang E (2021) Unbiased classification through bias-contrastive and bias-balanced learning. Adv Neural Inf Process Syst 34
117. Boedi LH, Grabner H (2021) Learning to ignore: fair and task independent representations. arXiv preprint arXiv:210104047
118. Wang M, Deng W (2020) Mitigating bias in face recognition using skewness-aware reinforcement learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020, vol 10
119. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller MA (2013) Playing Atari with deep reinforcement learning. arXiv:1312.5602
120. Kim MP, Ghorbani A, Zou J (2019) Multiaccuracy: black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 247–254
121. Adel T, Valera I, Ghahramani Z, Weller A (2019) One-network adversarial fairness. In: Thirty-third AAAI conference on artificial intelligence
122. Boedi LH, Grabner H (2021) Learning to ignore: fair and task independent representations. arXiv:2101.04047
123. Xu D, Du W, Wu X (2020) Removing disparate impact of differentially private stochastic gradient descent on model accuracy. arXiv preprint arXiv:200303699
124. Zhao H, Coston A, Adel T, Gordon GJ (2019) Conditional learning of fair representations. arXiv preprint arXiv:191007162
125. Cotter A, Jiang H, Gupta MR, Wang S, Narayan T, You S, Sridharan K (2019) Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. J Mach Learn Res 20(172):1–59
126. Ravuri S, Vinyals O (2019) Seeing is not necessarily believing: limitations of biggans for data augmentation. In: International conference on learning representations workshop LLD
127. Ravuri S, Vinyals O (2019) Classification accuracy score for conditional generative models. arXiv preprint arXiv:190510887
128. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:180911096
129. Frid-Adar M, Klang E, Amitai M, Goldberger J, Greenspan H (2018) Synthetic data augmentation using GAN for improved liver lesion classification. In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). IEEE, pp 289–293
130. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: European conference on computer vision. Springer, pp 740–755
131. Wang M, Deng W, Hu J, Tao X, Huang Y (2019) Racial faces in the wild: reducing racial bias by information maximization adaptation network. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 692–702
132. Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web, pp 1171–1180
133. Wang A, Russakovsky O (2021) Directional bias amplification. arXiv preprint arXiv:210212594
134. Besserve M, Mehrjou A, Sun R, Schölkopf B (2020) Counterfactuals uncover the modular structure of deep generative models. In: International conference on learning representations
135. Chalmers TC, Smith H Jr, Blackburn B, Silverman B, Schroeder B, Reitman D, Ambroz A (1981) A method for assessing the quality of a randomized control trial. Control Clin Trials 2(1):31–49
136. Qi J, Niu Y, Huang J, Zhang H (2020) Two causal principles for improving visual dialog. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10860–10869
137. Wang T, Isola P (2020) Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. PMLR, pp 9929–9939
138. Atzmon Y, Kreuk F, Shalit U, Chechik G (2020) A causal view of compositional zero-shot recognition. arXiv preprint arXiv:200614610
139. Mansour Y, Schain M (2014) Robust domain adaptation. Ann Math Artif Intell 71(4):365–380
140. Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D (2019) Invariant risk minimization. arXiv preprint arXiv:190702893
141. Creager E, Jacobsen JH, Zemel R (2021) Environment inference for invariant learning. In: International conference on machine learning. PMLR, pp 2189–2200
142. Adragna R, Creager E, Madras D, Zemel R (2020) Fairness and robustness in invariant learning: a case study in toxicity classification. arXiv preprint arXiv:201106485
143. Cao Y, Berend D, Tolmach P, Amit G, Levy M, Liu Y, Shabtai A, Elovici Y (2020) Out-of-distribution detection and generalization to enhance fairness in age prediction. arXiv preprint arXiv:200905283
144. Cheng P, Hao W, Yuan S, Si S, Carin L (2021) Fairfil: Contrastive neural debiasing method for pretrained text encoders. arXiv preprint arXiv:210306413
145. Van MH, Du W, Wu X, Lu A (2021) Poisoning attacks on fair machine learning. arXiv:2110.08932
146. Dwork C (2008) Differential privacy: a survey of results. In: International conference on theory and applications of models of computation. Springer, pp 1–19
147. Hébert-Johnson U, Kim MP, Reingold O, Rothblum GN (2017) Calibration for the (computationally-identifiable) masses. arXiv preprint arXiv:171108513
148. Ekstrand MD, Joshaghani R, Mehrpouyan H (2018) Privacy for all: ensuring fair and equitable privacy protections. In: Conference on fairness, accountability and transparency, pp 35–47
149. Xu D, Yuan S, Wu X (2019) Achieving differential privacy and fairness in logistic regression. In: Companion proceedings of The 2019 world wide web conference, pp 594–599
150. Hajian S, Domingo-Ferrer J, Monreale A, Pedreschi D, Giannotti F (2015) Discrimination-and privacy-aware patterns. Data Min Knowl Disc 29(6):1733–1782
151. Hajian S, Bonchi F, Castillo C (2016) Algorithmic bias: from discrimination discovery to fairness-aware data mining. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 2125–2126
152. Ruggieri S, Pedreschi D, Turini F (2010) Data mining for discrimination discovery. ACM Trans Knowl Discov Data (TKDD) 4(2):1–40
153. Kashid A, Kulkarni V, Patankar R (2017) Discrimination-aware data mining: a survey. Int J Data Sci 2(1):70–84
154. Bagdasaryan E, Poursaeed O, Shmatikov V (2019) Differential privacy has disparate impact on model accuracy. In: Advances in neural information processing systems, pp 15453–15462

155. Kocaoglu M, Snyder C, Dimakis AG, Vishwanath S (2018) Causalgan: Learning causal implicit generative models with adversarial training. arXiv:1709.02023

156. Zhang L, Wu Y, Wu X (2016) A causal framework for discovering and removing direct and indirect discrimination. arXiv preprint arXiv:161107509

157. Kim B, Wattenberg M, Gilmer J, Cai CJ, Wexler J, Viégas FB, Sayres R (2018) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: ICML

158. Cummings R, Gupta V, Kimpara D, Morgenstern JH (2019) On the compatibility of privacy and fairness. In: Adjunct publication of the 27th conference on user modeling, adaptation and personalization

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.