# A survey on bias in visual datasets

Simone Fabbrizzi [a,c,*], Symeon Papadopoulos [a], Eirini Ntoutsi [b,1], Ioannis Kompatsiaris [a]

[a] *CERTH-ITI, Thessaloniki, Greece*
[b] *Forschungsinstitut CODE, Universität der Bundeswehr, München, Germany*
[c] *Leibniz Universität, Hannover, Germany*

## ARTICLE INFO

## ABSTRACT

Computer Vision (CV) has achieved remarkable results, outperforming humans in several tasks. Nonetheless, it may result in significant discrimination if not handled properly. Indeed, CV systems highly depend on training datasets and can learn and amplify biases that such datasets may carry. Thus, the problem of understanding and discovering bias in visual datasets is of utmost importance; yet, it has not been studied in a systematic way to date. Hence, this work aims to: (i) describe the different kinds of bias that may manifest in visual datasets; (ii) review the literature on methods for bias discovery and quantification in visual datasets; (iii) discuss existing attempts to collect visual datasets in a bias-aware manner. A key conclusion of our study is that the problem of bias discovery and quantification in visual datasets is still open, and there is room for improvement in terms of both methods and the range of biases that can be addressed. Moreover, there is no such thing as a bias-free dataset, so scientists and practitioners must become aware of the biases in their datasets and make them explicit. To this end, we propose a checklist to spot different types of bias during visual dataset collection.

## 1. Introduction

In the fields of Artificial Intelligence (AI), algorithmic fairness, and (big) data ethics, the term *bias* usually refers to the case in which AI-powered decisions show prejudice against individuals or groups of people defined on protected attributes like gender or race (Ntoutsi et al., 2020). Instances of this prejudice have caused discrimination in many fields, including recidivism scoring (Angwin et al., 2016), online advertisement (Sweeney, 2013), facial recognition (Cook et al., 2019), and credit scoring (Bartlett et al., 2019).

Defining the concepts of bias and fairness in mathematical terms is not a trivial task. Verma and Rubin (2018) provided a survey on more than 20 different measures of algorithmic fairness, many of which are incompatible with each other. This incompatibility – the so-called impossibility theorem (Chouldechova, 2017; Kleinberg et al., 2017) – forces scientists and practitioners to choose the measures they use based on their personal beliefs or other constraints (e.g., business models) on what has to be considered fair for the particular problem/domain.

While algorithms may also be responsible for the amplification of pre-existing biases in the training data (Bolukbasi et al., 2016), issues in the quality of the data itself could contribute significantly to the development of discriminatory AI applications. Ntoutsi et al. (2020) identified two ways in which bias is encoded in the data: correlations and causal influences among the protected attributes and other features; and the lack of representation of protected groups in the data. They also noted that biases manifest in ways that are specific to the data type.

In this work, we focused on how biases can be encoded in the data (e.g., via spurious correlations, causal relationship among the variables, and unrepresentative data samples) and, in particular, in *visual* data (i.e., images and videos), which comprises one of the most popular and complex data types. Visual data encapsulates many features that require human experience and context to interpret. These include the human subjects, how they are depicted and their reciprocal position in the image frame, implicit references to cultural norms and background knowledge, etc. Even the colouring scheme can convey different messages. Thus, making sense of visual content remains a very complex task, and understanding bias in visual data is even harder.

Computer vision (CV), the primary domain that enables computers to gain high-level understanding from visual data, is heavily dominated nowadays by deep learning (DL) methods (LeCun et al., 2015; Baraniuk et al., 2020) that allowed for outstanding performance in tasks like object detection, image classification and image segmentation. DL methods, however, rely heavily on data, and the results are as good and "fair" as the data used for their training. CV has recently drawn attention for its ethical implications when deployed in several settings,

---

ranging from targeted advertising to law enforcement. There has been mounting evidence that deploying CV systems without comprehensive ethical assessment may result in major discrimination against protected groups. For instance, facial recognition technologies (Cook et al., 2019; Robinson et al., 2020), gender classification algorithms (Buolamwini and Gebru, 2018), and autonomous driving systems (Wilson et al., 2019) have been all shown to exhibit discriminatory behaviour.

While bias in AI systems is a well-studied field, the research in biased CV is more limited despite the abundance of visual data produced nowadays and their widespread use in the ML community. Moreover, to the best of our knowledge, there is no comprehensive survey on bias in visual datasets. Torralba and Efros (2011) represents a seminal work in the field, but it is limited to object detection datasets. Hence, the contributions of the present work are: (i) to explore and discuss the different types of bias that arise from the collection of visual data; (ii) to systematically review the works that aim at addressing and measuring bias in visual datasets; (iii) to discuss some attempts to compile bias-aware datasets. We believe this work to be a useful tool for helping scientists and practitioners to both develop new bias-discovery methods and collect data in ways as less biased as possible. To the latter end, we propose a checklist that can be used to spot the different types of bias that might enter the data during the collection process ( Table 6).

The structure of the survey is as follows. First, we describe in detail the different types of bias that might affect visual datasets (Section 2), provide concrete examples of CV applications that are affected by those biases, and a description of how they manifest in the life cycle of visual content. Second, we systematically review the methods for bias discovery in visual content proposed in the literature (Section 3). Third, in Section 4, we discuss the weaknesses and strengths of some bias-aware visual benchmark datasets. Finally, in Section 5, we conclude and outline some possible future research direction.

## 2. Manifestation of bias in visual data

In this section, we describe in detail the types of bias that pertain to the *capture* and *collection* of visual data (Fig. 1), namely: selection bias (Section 2.1), framing bias (Section 2.2) and label bias (Section 2.3). Furthermore, we describe (Section 2.4) how they manifest within the life cycle of visual content, from capture to the deployment of CV algorithms. Note that a comprehensive analysis of historical discrimination and algorithmic bias is beyond the scope of this work. The interested reader can refer to Bandy (2021) for a survey on methods for auditing algorithmic bias both in CV and other AI-related areas.

Our categorisation builds on the scheme by Torralba and Efros (2011) who organised the types of visual bias into four different categories: *selection bias*, *capture bias* (which we collapse into the more general concept of *framing bias*), *label bias*, and *negative set bias*. The latter arises when the labelling does not reflect entirely the population of the negative class (say *non-white* in a binary feature [white people/non-white people]). We consider negative class bias as an instance of selection and label bias.

Even though our categorisation appears on the surface to be similar to the one by Torralba and Efros (2011), their analysis focused on datasets for object detection. Instead, we contextualise bias in a more general setting and we also focus on discrimination against protected groups.[2] Since selection, framing and label bias manifest in many different ways, we also go further by describing a sub-categorisation of these three types of bias ( Table 1) including several biases commonly encountered in Statistics, Health studies, or Psychology and adapting them to the context of visual data. While in the following we describe selection, framing, and label bias in general terms, we also provide references in Table 1 for the interested reader who might want to delve further into their different manifestations.

### 2.1. Selection bias

*Definition.* Selection bias is the type of bias that "occurs when individuals or groups in a study differ systematically from the population of interest leading to a systematic error in an association or outcome".[3] More generally, it refers to any "association created as a result of the process by which individuals are selected into the analysis" (Hernán and Robins, 2020, Chapter 8, pg. 99). In visual datasets, using the first definition would be tricky as, for instance, in the case of facial recognition, respecting the ethnic composition of the population is generally not enough to ensure good performance across every subgroup, as we will see in the following. Hence, we adopt a slight modification of the definition by Hernán and Robins (2020):

> *We call* selection bias *any disparities or associations created as a result of the process by which subjects are included in a visual dataset.*

*Description.* Torralba and Efros (2011) showed that certain kinds of imagery are more likely to be selected during the collection of large-scale benchmark datasets, leading to selection bias (see *sampling bias*, Table 1). For example, in Caltech101 (Fei-Fei et al., 2004) pictures labelled as cars are usually taken from the side, while ImageNet (Deng et al., 2009) contains more racing cars. Furthermore, a strong selection bias can also be present within datasets. Indeed, Salakhutdinov et al. (2011) showed that, unless a great effort is made to keep the distribution uniform during the data collection process, categories in large image datasets follow a long-tail distribution. Having such a distribution means that, for example, people and windows are way more common than ziggurats and coffins (Zhu et al., 2014).

Selection bias becomes particularly worrisome when it concerns humans. Buolamwini and Gebru (2018) pointed out that under-representation of people from different genders and ethnic groups may result in a systematic misclassification of these groups (their work concentrates on gender classification algorithms). They also showed that some popular datasets were biased towards lighter-skinned male subjects. For example, Adience (Eidinger et al., 2014) and IJB-A (Klare et al., 2015) have, respectively, 7.4% and 4.4% of darker-skinned females; 6.4% and 16.0% of darker-skinned males; 44.6% and 20.2% of lighter-skinned females; and 41.6% and 59.4% of lighter-skinned males. Such imbalances greatly affect the performance of CV tools. For instance, Buolamwini and Gebru (2018) showed that the error rate for dark skin individuals could be 18 times higher than for light skin ones in some commercial gender classification algorithms.

*Affected applications.* In summer 2020, the New York Times published the story of a Black American individual wrongfully arrested due to an error made by a facial recognition algorithm (Hill, 2020). While we do not know whether bias in the data caused this exact case, we know that selection bias can lead to different error rates in face recognition. Hence, such technology would require much more care, especially in high-impact applications.

Autonomous driving systems are also likely affected by selection bias as it is very challenging to collect a dataset that describes every possible scene and situation a car might face. The Berkeley Driving Dataset (Yu et al., 2020) for example, contains driving scenes from only four cities in the US; an autonomous car trained on such a dataset will likely under-perform in other cities with different visual characteristics. The effect of selection bias on autonomous driving becomes particularly risky when it affects pedestrian recognition algorithms. Wilson et al. (2019) studied the impact of under-representation of darker-skinned people on the predictive inequity of pedestrian recognition systems.

---

[2] Note that, while this is the focus of our survey, we also take into account cases in which bias does not necessarily affect people, e.g., in object detection.
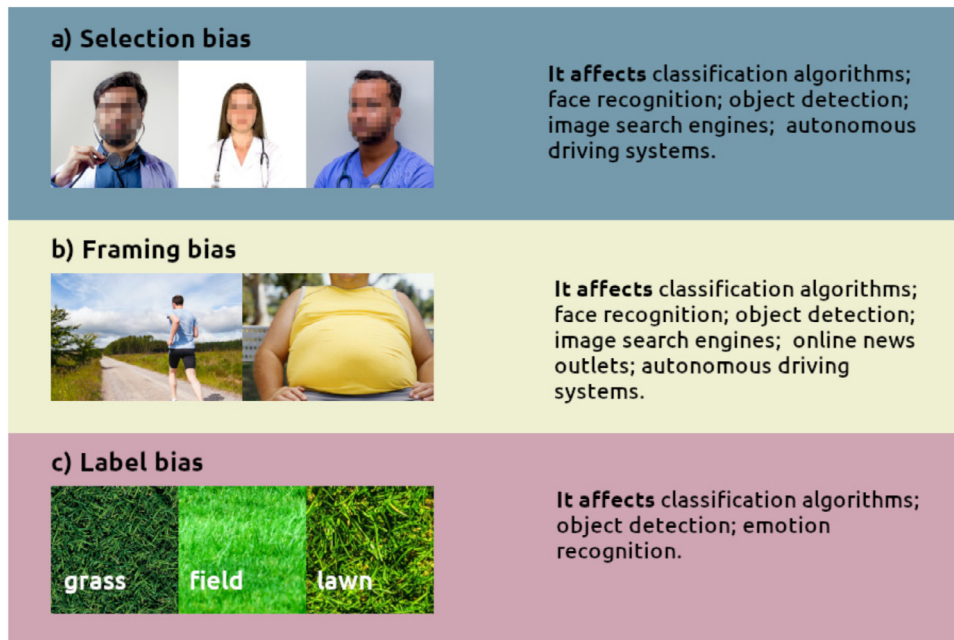
[3] Catalogue of Bias Collaboration, Nunan D, Bankhead C, Aronson JK. Selection bias. Catalogue Of Bias 2017: http://www.catalogofbias.org/biases/selection-bias/. Last visited 18.05.2022.

**Fig. 1.** Examples of selection, framing and label bias. On the right, a list of applications that can be affected by each type of bias.

They found evidence that the effect of this selection bias is two-fold: first, such imbalances "beget less statistical certainty" making the process of recognition more difficult; second, standard loss functions tend to prioritise the more represented groups and hence some kind of mitigation measures are needed in the training procedure (Wilson et al., 2019).

Moreover, Buolamwini and Gebru (2018) explained that part of the collection of benchmark face datasets is often done using facial detection algorithms. Therefore, every systematic bias in the training of those tools propagates to other datasets. This is a clear example of *algorithmic bias turning into selection bias* (see *automation bias*, Table 1), as described at the end of Section 2.4. Furthermore, image search engines can contribute to the creation of selection biases (see *availability bias*, Table 1) as well due to systematic biases in their retrieval algorithms (see, Kay et al., 2015) for a study on gender bias in Google's image search engine.

*Remarks.* Finally, Klare et al. (2012) pointed out that, while class imbalances have undoubtedly significant impact on some facial recognition algorithms, they do not explain every disparity in the performance of algorithms. For instance, they suggested that a group of subjects might be more challenging to recognise, even with balanced training data, if it is associated with higher variance (for example, due to hairstyle or make-up). The reader can also refer to Terhörst et al. (2021b) for an analysis of the effect of these kinds of attributes on facial recognition technology.

### 2.2. Framing bias

*Definition.* According to the seminal work of Entman (1993) on framing in (textual) communication, "To frame is to *select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation and/or treatment recommendation* for the item described". Coleman (2010) adopted the same definition for the framing of visual content and added that "In visual studies, *framing* refers to the selection of one view, scene, or angle when making the image, cropping, editing or selecting it". These two definitions highlight how framing bias has two different aspects. First, the medium aspect: visual content is, in all respect, a medium, and therefore the

way a visual scene is composed conveys different messages. Second, the technical aspect: framing bias also derives from how an image is captured or edited. Hence, in the following:

> *We refer to framing bias as any associations or disparities that can be used to convey different messages and/or that can be traced back to the way in which the visual content has been composed.*

Note that, while the selection process is a powerful tool for framing visual content, we keep the concepts of selection and framing bias distinct, as they present their own peculiarities.

*Description.* An example of visual framing bias (to be more specific, *capture bias*, Table 1) has been studied by Heuer et al. (2011). They analysed images attached to obesity-related articles that appeared on major US online news websites in the period 2002–2009. They concluded that there was a substantial difference in how such images depicted obese people with respect to non-overweight ones. For example, 59% of obese people were headless (6% of non-overweight), and 52% had only the abdomen portrayed (0% of non-overweight). This portrayal as "headless stomachs" (see Fig. 1 (b) for an example) may have a stigmatising and de-humanising effect on the viewer (Heuer et al., 2011). On the contrary, positive characteristics were more commonly portrayed among non-overweight people. Corradi (2012), while analysing the use of female bodies in advertisements, talked about "semiotic mutilation" when parts of a woman's body are used for advertising a product with a de-humanising effect, similar to what Heuer et al. have described in their work about obesity.

The relationship between bodies and faces in image framing is a well-known problem. Indeed, Archer et al. (1983) used a ratio between the height of a subject's body and the length of their face to determine whether men's faces were more prominent than those of women. They found out that this was true in three different settings: contemporary American news media photographs, contemporary international news media photographs, and portraits and self-portraits from the 17-th century to the 20-th century (interestingly, there was no disparity in earlier artworks). Furthermore, they found some evidence that people tend to draw men with higher facial prominence and thus that this bias does not only occur in mass media or art.

*Affected applications.* An application that can suffer greatly from framing bias (see *stereotyping*, Table 1) is that of image search engines. For example, Kay et al. (2015) found out that while searching for construction workers on Google Image Search, women were more likely to be depicted in an unprofessional or hyper-sexualised way.[4] However, it is unclear whether the retrieval algorithms are responsible for the framing or they just index popular pages associated with the queries. Nonetheless, the problem remains because the biased framing in a search engine can contribute to the spread of harmful or misleading messages. We recall, for instance, the case of the photo of a woman "ostensibly about to remove a man's head" (Paquette, 2015) that was retrieved after searching "feminism" on a famous stock photos website.

### 2.3. Label bias

*Definition.* For supervised learning, labelled data are required. The quality of the labels[5] is of paramount importance for learning and comprises a tedious task due to the complexity and volume of today's datasets. Jiang and Nachum (2020) define label bias as "*the bias that arises when the labels differ systematically from the ground truth*". For instance, Terhörst et al. (2021a) found some evidence that the face recognition dataset Labelled Faces in the Wild (Huang et al., 2008; Kumar et al., 2009) presents a very low label accuracy against human annotation.

Furthermore, Torralba and Efros (2011) highlight bias as a result of the labelling process itself for reasons like "semantic categories are often poorly defined, and different annotators may assign different labels to the same type of object". Torralba and Efros' work mainly focused on object detection tasks, and hence by *different label assignment* they refer, e.g., to "grass" labelled as "lawn" or "picture" as "painting". Yet, these problems arise especially when dealing with human-related features such as race or gender.

> We define label bias as any errors in the labelling of visual data, with respect to some ground truth, or the use of poorly defined or inappropriate semantic categories.

*Description.* As already mentioned, a significant source of labelling bias is the *poor definition of semantic categories*. Race is a particularly clear example of this: according to Barbujani and Colonna (2011) "The obvious biological differences among humans allow one to make educated guesses about an unknown person's ancestry, but agreeing on a catalogue of human races has so far proved impossible". Given such an impossibility, racial categorisation in visual datasets must come at best from subjects' own race perception or, even worse, from the stereotypical bias of annotators. From a CV standpoint then, it would be probably more accurate to use actual visual attributes, if strictly necessary, such as the *Fitzpatrick skin type* (Buolamwini and Gebru, 2018) or *skin reflectance* (Cook et al., 2019) rather than a fuzzy category such as race. Note that, while skin tone can be a more objective trait, it still does not entirely reflect human diversity. Similarly, the binary categorisation of gender has been criticised. The reader can refer to Hanna et al. (2020), Kasirzadeh and Smart (2021) to explore the challenges that the use of ill-defined categories poses to algorithmic fairness from both the ontological and operational points of view.

Moreover, Torralba and Efros (2011) argued that different annotators can come up with different labels for the same object. While this mainly applies to the labelling of object detection datasets rather than face datasets, where labels are usually binary or discrete, it gives us

an important input about bias in CV in general: annotators' biases and preconceptions are reflected in the datasets.

We can view what has been described so far also as a problem of operationalisation of what Jacobs and Wallach (2021) called unobservable theoretical constructs (see *measurement bias*, Table 1). They proposed a framework that serves as guideline for the mindful use of fuzzy semantic categories and answers the following questions on the validity of operationalisations (or measurements) of a construct: "Does the operationalisation capture all relevant aspects of the construct purported to be measured? Do the measurements look plausible? Do they correlate with other measurements of the same construct? Or do they vary in ways that suggest that the operationalisation may be inadvertently capturing aspects of other constructs? Are the measurements predictive of measurements of any relevant observable properties (and other unobservable theoretical constructs) thought to be related to the construct, but not incorporated into the operationalisation? Do the measurements support known hypotheses about the construct? What are the consequences of using the measurements?"

A concrete case of the use of a fuzzy semantic category is provided in Liang et al. (2018), where the creation of a face dataset for facial beauty is described. Since beauty and attractiveness are prototypical examples of subjective characteristics, it is obvious that attempts of constructing such datasets will be filled with the personal preconceptions of the participants who labelled the images (see *observer bias*, Table 1).

*Affected applications.* Since deep learning boosted the popularity of CV, a modern form of physiognomy has gained a certain momentum. Recently, several studies appeared claiming to be able to classify images according to the criminal attitude[6] or sexual orientation[7] of the subjects. A commercial tool has also been released to detect terrorists and paedophiles. While "doomed to fail" for a series of technical reasons well explained by Bowyer et al. (2020), these applications rely on a precise ideology that Goldenfein (2019) called *computational empiricism*: an epistemological paradigm that claims, despite any scientific evidence, that the true nature of humans can be measured and unveiled by algorithms. The reader can also refer to the famous blog post "Physiognomy's New Clothes"[8] for an introduction to the problem.

*Remarks.* Just as selection bias, label bias can lead to a vicious cycle: a classification algorithm trained on biased labels will most likely reinforce the original bias when used to label newly collected data (see *automation bias*, Table 1).

### 2.4. Media bias life cycle

Fig. 2 gives an overview of the life cycle of visual content and depicts how different types of bias can enter at any step of this cycle and can be amplified in consecutive interactions. Below we describe each of these components in more detail.

*1. Real world.* The journey of visual content alongside bias starts even before the content is generated. Inequalities undeniably shape our world, and this is reflected in the generation of data in general and of visual content in particular. For example, Zhao et al. (2017) found out that the dataset MS-COCO (Lin et al., 2014), a large-scale object detection, segmentation, and captioning dataset which is used as a benchmark in CV competitions, was more likely to associate kitchen

---

[4] This specific problem seems to be solved, but the queries `female construction worker` and `male construction worker` still return several pictures in which the subjects are hyper-sexualised.

[5] Note that by label we mean any tabular information attached to the image data (object classes, measures, protected attributes, etc.).

[6] See, for instance, the retracted article: Hashemi and Hall https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0282-4. Last visited 30.03.2022.

[7] For example, Kosinski and Wang https://www.gsb.stanford.edu/faculty-research/publications/deep-neural-networks-are-more-accurate-humans-detecting-sexual. Last visited 30.03.2022.

[8] B. Agüera y Arcas, M. Mitchell and A. Todorov, *Physiognomy's New Clothes*, May 2017. https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a. Last visited 30.03.2022.

**Table 1**
A finer-grained categorisation of biases described in Section 2.

| Name | Description | Selection | Framing | Label |
|---|---|:---:|:---:|:---:|
| Sampling bias[b] | Bias that arises from the sampling of the visual data. It includes class imbalance. | • | | |
| Negative set bias (Torralba and Efros, 2011) | When a negative class (say non-white in a white/non-white categorisation) is not representative enough. | • | | • |
| Availability bias[a] | Distortion arising from the use of the most readily available data (e.g., using search engines). | • | | |
| Platform bias | Bias that arises as a result of a data collection being carried out on a specific digital platform (e.g., Twitter, Instagram, etc.). | • | | |
| Volunteer bias[a] | When data is collected in a controlled setting instead of being collected in-the-wild, volunteers that participate in the data collection procedure may differ from the general population. | • | | |
| Crawling bias | Bias that arises as a result of the crawling algorithm/system used to collect images from the Web or with the use of an API (e.g., the keywords used to query an API, the seed websites used in a crawler). | • | • | |
| Spurious correlation | Presence of spurious correlations in the dataset that falsely associate a certain group of subjects with any other features. | • | • | |
| Exclusion bias[b] | Bias that arise when the data collection excludes partly or completely a certain group of people. | • | • | |
| Chronological bias[a] | Distortion due to temporal changes in the visual world the data is supposed to represent. | • | • | • |
| Geographical bias (Shankar et al., 2017) | Bias due to the geographic provenance of the visual content or of the photographer/video maker (e.g., brides and grooms depicted only in western clothes). | • | • | |
| Capture bias (Torralba and Efros, 2011) | Bias that arise from the way a picture or video is captured (e.g., objects always in the centre ,exposure, etc.). | | • | |
| Apprehension bias[a] | Different behaviour of the subjects when they are aware of being photographed/filmed (e.g., smiling). | | • | |
| Contextual bias (Singh et al., 2020) | Association between a group of subjects and a specific visual context (e.g., women and men respectively in household and working contexts) | | • | |
| Stereotyping[c] | When a group is depicted according to stereotypes (e.g., female nurses vs. male surgeons). | | • | |
| Measurement bias (Jacobs and Wallach, 2021) | Every distortion generated by the operationalisation of an unobservable theoretical construct (e.g., race operationalised as a measure of skin colour). | | | • |
| Observer bias[a] | Bias due to the way a annotator records the information. | | | • |
| Perception bias[a] | When data is labelled according to the possibly flawed perception of a annotator (e.g., perceived gender or race) or when the annotation protocol is not specific enough or is misinterpreted. | | | • |
| Automation bias[c] | Bias that arises when the labelling/data selection process relies excessively on (biased) automated systems. | • | | • |

The bullets represent which types of bias they are a subcategory of (Note that there are overlaps as different kind of bias can co-occur).

[a]The definitions are adapted from: https://catalogofbias.org/biases/.

[b]The definitions are adapted from: https://en.wikipedia.org/wiki/Sampling_bias.

[c]The definitions are adapted from: https://en.wikipedia.org/wiki/List_of_cognitive_biases; and Torralba and Efros (2011), Shankar et al. (2017), Singh et al. (2020) and Jacobs and Wallach (2021).

objects with women. While both image capturing and dataset collection come at a later stage in the life cycle described in Fig. 2, it is clear that in this instance, such bias has roots in the gender division between productive and reproductive/care labour. Nevertheless, as shown in the following paragraphs, each step of the life cycle of visual content can reproduce or amplify historical discrimination as well as insert new biases.

*2. Capture.* The actual life of visual content starts with its capture. Here the first types of bias can be introduced, *selection bias* and *framing bias*. While selection bias is usually observed in datasets, where entire groups can be under-represented or non-represented at all, the selection begins with the choices of the (professional or amateur) photographer/video maker. Moreover, the way a photo is composed (arrangement, camera's setting, lighting conditions, etc.) is a powerful way of conveying different messages and thus a possible source of framing bias. Note that the selection of the subjects and the framing of the photo/video are both active choices of the photographer/video maker. Nevertheless, historical discrimination can turn into selection and framing bias if not actively countered.

*3. Editing.* With the advent of digital photography, image editing and post-processing are now a key step in the content life cycle. Post-processing has become a fundamental skill for every photographer/video maker, along with skills such as camera configuration, lighting, shooting, etc. Since photo editing tools are extremely powerful, they could give rise to many ethical issues: to what extent and in which contexts is it right to modify the visual content of an image or video digitally? What harms could such modifications potentially cause? The discussion around the award-winning Paul Hansen's photo "Gaza Burial" (Krug and Niggemeier, 2013) represents a practical example of how these questions are important to photojournalism. In particular, what is the trade-off between effective storytelling and adherence to reality? Nonetheless, photo editing does not concern journalism only. Actually, it affects people, and especially women, in several different contexts, from the fashion industry (Jamil, 2018) to advertising and high-school yearbooks (Cramer and Levenson, 2021).

*4. Dissemination.* The next step in the life of visual content is dissemination. Nowadays, images and videos are shared via social and mainstream media in volumes that nobody can possibly inspect. For instance, more than 500 h of videos are uploaded to YouTube every minute.[9] Dissemination of visual content suffers from both selection

---

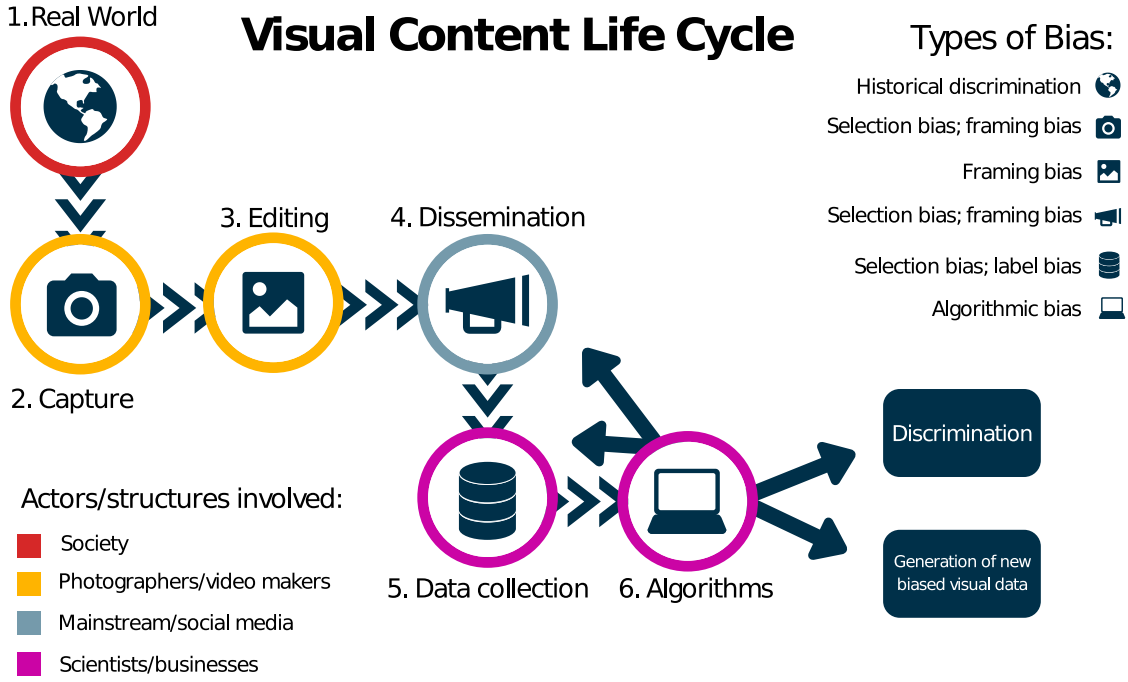[9] https://blog.youtube/press/ Last visited 23.03.2022.

**Fig. 2.** Simplified illustration of visual content life cycle and associated sources of bias.

and framing bias. The images that are selected, the medium and channels through which they are disseminated, the caption or text that are attached to them are all elements that could give rise to bias. See, for example, Peng (2018) for a case study of framing bias in the media coverage of 2016 US presidential election.

*5. Data collection.* The part of the life cycle of visual content that is more relevant to our discussion is the collection of visual datasets. Here we encounter selection bias once again, as the data collection process can exclude or under-represent certain groups from appearing in the dataset, as well as *label bias* as great effort is usually expended to collect data along with additional information in the form of annotations. As we discussed in Section 2.3, this process is prone to errors, mislabelling, and explicit discrimination. See Miceli et al. (2020) for an interesting analysis of power dynamics in the labelling process of visual data. Furthermore, benchmark datasets have gained incredible importance in the CV field. On the one hand, they allowed significant measurable progress in the field. On the other hand, they represent only a small portion of the visual world. The reader can refer to Prabhu and Birhane (2020) for a critique of the large-scale collection of visual data *in the wild*.

*6. Algorithms.* Finally, there is the actual step of model training. Fairness and accountability of algorithms is a pressing issue as algorithm-powered systems are used pervasively in applications and services impacting a growing number of citizens (Drozdowski et al., 2020; Bandy, 2021). Important questions arise for the AI and CV communities on how to implement fairness in algorithms such as: What legal frameworks should governments put into place? What are the strategies to mitigate bias or make algorithms explainable? Nevertheless, the journey of visual content does not end with the training of models. Indeed, there are several ways in which biased algorithms can generate vicious feedback loops as described in the remarks of Section 2.3. Moreover, the recent explosion in popularity of generative models, namely Generative Adversarial Networks (Goodfellow et al., 2014), has made the process of media creation very easy and fast. See Mirsky and Lee (2021) for a survey on AI-generated visual content. Such AI-generated media can then be reinserted in the content life cycle via the Web and present their own ethical issues. Refer to Berendt et al. (2021) for a discussion on the socio-technical aspects and ambivalence of the Web of humans and AIs.

**Table 2**
Brief summary of the notation used in Section 3.

| Symbol | Description |
| --- | --- |
| $\mathcal{D}$ | Cursive capital letters denotes visual datasets |
| $f_D(\cdot)$ | A model $f$ trained on the dataset $\mathcal{D}$ |
| $\mathbf{w}$ | Bold letters denotes vectors |
| $AP$ | Average Precision-Recall (AP) score |
| $AP_B(f_D)$ | AP score of the model $f_D$ when tested on the dataset $\mathcal{B}$ |
| $\|\cdot\|_2$ | L2-norm |
| $|\mathcal{D}|$ | The number of elements in the dataset $\mathcal{D}$ |
| $\sigma(\cdot)$ | Sigmoid function |
| $\mathbf{1}[\cdot]$ | Indicator function |
| $\mathbb{P}(\cdot)$ | Probability |
| $\mathbb{P}(\cdot|\cdot)$ | Conditional probability |
| $H(\cdot)$ | Shannon entropy |
| $H(\cdot|\cdot)$ | Conditional Shannon entropy |
| $I(\cdot,\cdot)$ | Mutual information |
| $D(\cdot)$ | Simpson $D$ score |
| $ln(\cdot)$ | Natural logarithm |
| $mean(\cdot)$ | Arithmetic mean |

## 3. Bias discovery and quantification in visual datasets

This section aims to understand how researchers have tackled the problem of discovery and quantification of bias in visual datasets since high-quality visual datasets are a critical ingredient towards fair and trustworthy CV systems (Hu et al., 2020). To this end, we performed a systematic survey of papers addressing the following problem: *Given a visual dataset $\mathcal{D}$, is it possible to discover/quantify what types of bias it exhibits (of those described in Section 2)?* In particular, we focus on the methodologies and measures used in the bias discovery process, of which we are going to outline the pros and cons. Furthermore, we will try to define open issues and possible future directions for research in this field. Note that this problem is critically different from both the problem of finding out whether an algorithm discriminates against a protected group and the problem of mitigating such bias. In order to systematise this review, we proceed in the following way to collect the relevant material: first, we outlined a set of keywords to be used in three different scientific databases (DBLP, arXiv and Google Scholar) and we selected only the material relevant to our research question

**Table 3**

Set of keywords and relative synonyms and antonyms.

|  | Synonyms | Antonyms |
|---|---|---|
| bias | discrimination | fair, unbiased |
| image | visual, face, facial |  |
| dataset | collection |  |

following a protocol described in the following paragraph; second, we summarised the results of our review and outlined the pros and cons of different methods in Section 3.5. The reader can refer to Table 2 for a summary of the mathematical notation used. Our review methodology was inspired by the works of Merli et al. (2018), and Kofod-Petersen (2012). Our protocol also resembles the one described in Kitchenham (2004, Section 5.1.1).

Given our problem, we identify the following relevant keywords: *bias, image, dataset, fairness*. For each of them, we also defined a set of synonyms and antonyms (see Table 3). Note that we include the words "face" and "facial" among the synonyms of the word image. This is mainly motivated by the fact that in the title and the abstract of the influential work of Buolamwini and Gebru (2018) there are no occurrences of the word "image". Instead, we find many occurrences of the expression "facial analysis dataset". Since facial analysis is an important case study for detecting bias in visual content, it makes sense to include it explicitly in our search. The search queries have been composed of all possible combinations of the different synonyms (antonyms) of the above keywords (for example, "image dataset bias", "visual dataset discrimination", or "image collection fairness").

This process resulted, after manual filtering,[10] in 17 relevant papers. We expanded the list with 6 articles by looking at papers citing the retrieved works (via Google Scholar) and their "related work" sections. We also added to the list the work of Lopez-Paz et al. (2017) on causal signals in images that was not retrieved by the protocol described above. In the following, we review all these 24 papers (see Table 4), dividing them into four categories according to the strategies they use to discover bias:

- **Reduction to tabular data:** these rely on the attributes and labels attached to or extracted from the visual data and try to measure bias as if it were a tabular dataset.
- **Biased image representations:** these rely on lower-dimensional representations of the data to discover bias.
- **Cross-dataset bias detection:** these assess bias by comparing different datasets, trying to discover some sort of "signature" due to the data collection process.
- **Other methods:** Different methods that could not fit any of the above categories.

### 3.1. Reduction to tabular data

Methods in this category transform visual data into tabular format and leverage the multitude of bias detection techniques developed for tabular datasets (Ntoutsi et al., 2020). The features for the tabular description can be extracted either *directly* from the images (using, for example, some image recognition tools) or *indirectly* from some accompanying image description/annotation (e.g., image captions, labels/hashtags) or both. In the majority of cases, feature extraction relies on automatic processes and is therefore prone to errors and biases. Thus, biases that exist in the original images (selection, framing,

---

[10] The manual filtering consisted in keeping only those papers that describe a method or a measure for discovering and quantifying bias in visual datasets. In some cases, we kept also works developing methods for algorithmic bias mitigation but that could be used to discover bias in the dataset as well. E.g., Balakrishnan et al. (2020).

or label bias) might be reflected and even amplified in the tabular representation due to the bias-prone feature extraction process.

Below we review different approaches under this category that extract protected attributes from visual datasets and evaluate whether bias exists. Existing approaches can be broadly categorised into parity-based, information theoretical and others. As already explained, the sources of bias in such a case are not limited to bias in the original images, but bias may also exist due to the labelling process and the automatic feature extraction. The impact of such additional sources on the results is typically omitted.

*Count/demographic parity.* Dulhanty and Wong (2019) proposed a simple method for auditing ImageNet (Deng et al., 2009) with respect to gender and age biases. They applied a face detection algorithm to two different subsets of ImageNet, the training set of ILSVRC (Russakovsky et al., 2015) and the *person* category of ImageNet (Deng et al., 2009). After that, they applied an age recognition model and a gender recognition model to them. They then computed the dataset distribution across the age and gender categories, finding out a prevalence of men (58.48%) and a minimal amount of people (1.71%) in the > 60 age group. Computing the percentage of men and women across every category allowed them to identify the highest class imbalances in the two subsets of ImageNet. For example, in the ILSVRC subset, the 89.33% of the images in category *bulletproof vest* were labelled as men and the 82.81% of the images in category *lipstick* were labelled as women. Therefore, this method does not only give information on the selection bias but also on the *framing* of the protected attributes, given a suitable labelling of the dataset. The authors noted that this method relies on the assumption that the gender and age recognition models involved are not biased. As the authors pointed out, such an assumption is violated by the gender recognition model, and therefore, analysis cannot be totally reliable.

Yang et al. (2020) performed another analysis of the *person* category of ImageNet (Deng et al., 2009), trying to address both selection and label bias. They addressed label bias by asking annotators to find out, first, whether the labels could be offencive or sensitive (e.g., sexual/racial slur), and, second, to point out whether some of the labels were not referring to visual categories (e.g., it is difficult to understand whether an image depicts a *philanthropist*). They removed such categories and continued their analysis by asking annotators to further label images according to some categories of interest (gender, age, and skin colour) to understand whether the remaining data were balanced with respect to those categories and then to address selection bias. This demographic analysis showed that women and dark-skinned people were both under-represented in the remaining non-offencive/sensitive and visual categories. Moreover, despite the overall under-representation, some categories were found to align with stereotypes (e.g., the 66.4% of people in the category *rapper* were dark-skinned). Hence, they also potentially addressed some framing bias. The annotation process was validated by measuring the annotators' agreement on a small controlled set of images.

Zhao et al. (2017) measured the correlation between a protected attribute and the occurrences of various objects/actions. They assumed to have a dataset labelled with a protected attribute $G$ with values $\{g_1, \ldots, g_n\}$ and an attribute $O = o$ describing the occurrences of said objects or actions in the images (for instance, $G$ could be the gender and $O$ a cooking tool or an outdoor sport scene). The bias score of the dataset with respect to a protected attribute value $G = g$ and the scene/object $O = o$ is defined as the percentage of $g$ co-occurrences with $o$:

$$b(o, g) = \frac{c(o, g)}{\sum_{x \in \{g_1, \ldots, g_n\}} c(o, x)} \tag{1}$$

where $c(o, x)$ counts the co-occurrences of the object/scene $o$ and the protected attribute's value $x$. If $b(o, g_i) > \frac{1}{n}$, where $n$ is the number of values that the protected attribute can assume, it means that attribute $G = g$ is positively correlated with the object/action $O = o$.

**Table 4**
Summary of the collected material.

| | No. | Paper | Year | Selection | Framing | Label | Type of measures/methods |
|---|---|---|---|---|---|---|---|
| Reduction to tabular data | 1 | Dulhanty and Wong (2019) | 2019 | • | • | | Count; Demographic parity |
| | 2 | Yang et al. (2020) | 2020 | • | • | • | Count; Demographic parity |
| | 3 | Zhao et al. (2017) | 2017 | • | • | | Demographic parity |
| | 4 | Shankar et al. (2017) | 2017 | • | | | Count |
| | 5 | Buolamwini and Gebru (2018) | 2018 | • | | | Count |
| | 6 | Merler et al. (2019) | 2019 | • | | | Entropy-based; Information theoretical |
| | 7 | Panda et al. (2018) | 2018 | • | • | | Entropy-based |
| | 8 | Kim et al. (2019) | 2019 | • | • | | Information theoretical |
| | 9 | Wang et al. (2019) | 2019 | • | • | | Dataset leakage |
| | 10 | Wachinger et al. (2021) | 2021 | • | | | Causality |
| | 11 | Jang et al. (2019) | 2019 | | • | | 4 different measures |
| | 12 | Wang et al. (2020) | 2020 | • | • | • | 13 different measures |
| Biased image representation | 13 | Kärkkäinen and Joo (2021) | 2021 | • | | | Distance-based |
| | 14 | Steed and Caliskan (2021) | 2021 | | • | | Distance-based |
| | 15 | Balakrishnan et al. (2020) | 2020 | | • | | Interventions |
| Cross-dataset bias detection | 16 | Torralba and Efros (2011) | 2011 | • | • | | Cross-dataset generalisation |
| | 17 | Tommasi et al. (2015) | 2015 | • | • | | Cross-dataset generalisation |
| | 18 | Khosla et al. (2012) | 2012 | • | • | | Modelling bias |
| | 19 | López-López et al. (2019) | 2019 | • | | | Nearest neighbour in a latent space |
| Other | 20 | Model and Shamir (2015) | 2015 | • | • | | Model-based |
| | 21 | Thomas and Kovashka (2019) | 2019 | | • | | Model-based |
| | 22 | Clark et al. (2020) | 2020 | • | • | | Modelling bias |
| | 23 | Lopez-Paz et al. (2017) | 2017 | • | | | Causality |
| | 24 | Hu et al. (2020) | 2020 | • | • | | Crowd-sourcing |

Shankar et al. (2017) used instead a simple count to assess geographic bias in ImageNet (Russakovsky et al., 2015) and Open Images (Krasin et al., 2016). They found out, for example, that the great majority of images of which the geographic provenance is known comes from the USA or Western European countries, resulting in highly imbalanced data. Such a geography-based analysis has been expanded by Wang et al. (2020). While having balanced data is important for many applications, a mere count is usually not enough to assess every type of bias, as a balanced dataset could still contain spurious correlations, framing bias, label bias, etc.

Buolamwini and Gebru (2018) constructed a benchmark dataset for gender classification. For testing discrimination in gender classification models, their dataset is balanced according to the distribution of both gender and Fitzpatrick skin type as they noticed that the error rates of classification models tended to be higher at the intersection of those categories (e.g., black women) because of the use of imbalanced training data. Hence, while they quantify bias by simply counting the instances with certain protected attributes, the novelty of their work is that they took into account multiple protected attributes at a time.

*Information theoretical.* Merler et al. (2019) introduced four measures to construct a balanced dataset of faces. Two measures of diversity: Shannon entropy $H(X) = -\sum_{i=1}^{n} \mathbb{P}(X = x_i) \cdot ln(\mathbb{P}(X = x_i))$ and Simpson index $D(X) = \frac{1}{\sum_{i=1}^{n} \mathbb{P}(X=x_i)^2}$ where $\mathbb{P}(X = x_i)$ is the probability of an image having value $x_i$ for the attribute $X \in \{x_i\}_{i=1}^{n}$; and two measures of evenness: ($E_{Shannon} = \frac{H(X)}{ln(n)}$ and $E_{Simpson} = \frac{D(X)}{n}$). Such measures have been applied to a set of facial attributes, ranging from craniofacial distances to gender and skin colour, computed both via automated systems and with the help of human annotators.

Panda et al. (2018) also proposed to use (conditional) Shannon entropy for discovering *framing bias* in emotion recognition datasets. Using a pre-trained model, they computed the dataset's top occurring objects/scenes. They computed the conditional entropy of each object across the positive and negative set of the emotions to see whether some objects/scenes were more likely to be related to a certain emotion. For example, they found out that objects like balloons or candy stores are only present in the negative-set of *sadness* in Deep Emotion (You et al., 2016). Given an object $c$ and an emotion $E = e \in \{0, 1\}$ (where $e = 1$ represents, for instance, *sadness* and $e = 0$ represents the negative-set *non-sadness*) they computed:

$$H(E|X = c) = -\sum_{e \in \{0,1\}} \mathbb{P}(E = e|X = c) \cdot ln(\mathbb{P}(E = e|X = c)) \quad (2)$$

When the conditional entropy of an object is zero, it means that such an object is associated only with the emotion $E$ or, on the contrary, is never associated with it (it only appears in the negative set). This may be considered a type of framing bias.

Kim et al. (2019) introduced another definition of bias inspired by information theory. They wanted to develop a method for training classification models that do not overfit due to dataset biases. In doing so, they give a precise definition of bias: a dataset contains bias when $I(X, Y) \gg 0$ where $X$ is the protected attribute, $Y$ is the output variable and $I(X, Y) := H(X) - H(X|Y)$ is the mutual information between those random variables. Kim et al. proposed to minimise such mutual information during training so that the model forgets the biases and generalises well. Note that if $Y$ is any feature in the dataset, this measure can be used to quantify bias in the dataset instead of the output of a model.

*Other.* Wang et al. (2019) defined *dataset leakage* to measure the possibility of a protected attribute to be retrieved using only the information about non-protected ones. Given a dataset $\mathcal{D} = \{(x_i, y_i, g_i)\}$ where $x_i$ is an image, $y_i$ a non-protected attribute and $g_i$ is a protected attribute, the attribute $y_i$ is said to leak information about $g_i$ if there exists a function $f(\cdot)$, called *attacker*, such that $f(y_i) \approx g_i$. Operationally, the attacker $f(\cdot)$ is a classifier trained on $\{y_i, g_i\}$. The dataset leakage is measured as follows:

$$\lambda_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \mathbf{1}[f(y_i) = g_i] \quad (3)$$

Wachinger et al. (2021) explicitly used causality for studying spurious correlations in neuroimaging datasets. Given variables $X$ and $Y$ they wanted to test whether $X$ causes $Y$ or there exists a confounder variable $Z$ instead. Since those two hypotheses imply two different factorisations of the distribution $\mathbb{P}(X|Y)$, the factorisation with a lower Kolmogorov complexity is the one that identifies the true causal structure. Kolmogorov complexity is approximated by Minimum Description Length (MDL).

Jang et al. (2019) proposed 8 different measures for identifying gender framing bias in movies. The following measures are computed for a movie for each gender, therefore the means are computed across every frame in which an actor of a certain gender appears. The measures are the following: Emotional Diversity $H(X) = -\sum_{i=1}^{s} \mathbb{P}(X = x_i) \cdot ln(\mathbb{P}(X = x_i))$ where $\mathbb{P}(X = x_i)$ is the probability that a character

expresses a certain emotion $x_i$ and the sum is computed on the range of the different emotions shown by characters of a certain gender (the list of emotions was: anger, disgust, contempt, fear, happiness, neutral, sadness, and surprise); Spatial Staticity exploits ideas from time-series analysis to measure how static a character is during the movie, it is defined as $mean(PSD(p(t)))$ where PSD is the power spectral density of the time-series of the position on the $x$-axis (resp. y-axis) of the character (the higher the value the less animated is the character); Spatial occupancy $mean(\sqrt{A})$ where $A$ is the area of the face of the character; Temporal occupancy $\frac{N}{N_{total}}$ where $N$ is the number of frames in which the character appears and $N_{total}$ is the total number of frames; Mean age computed over each frame and character; Intellectual Image $mean(G)$ where $G$ is the presence of glasses (this seems a debatable choice as it might itself suffer from some label bias); Emphasis on appearance $mean(E)$ where $E$ is the light exposure of faces again calculated for each frame; finally, the type and frequency of surrounding objects is analysed. The attributes involved in the computation of these measures are mainly extracted using Microsoft Face API, which (Buolamwini and Gebru, 2018) demonstrated to be biased against black women.

Wang et al. (2020) proposed REVISE, a comprehensive tool for bias discovery in image datasets. The authors defined three sets of metrics. The first set contains metrics based solely on the use of bounding boxes of objects (Note that a *person* is considered an *object* as it can be classified by an objected detection model); the second set contains metrics that exploit also protected attribute labels; the third set uses additional unstructured information such as text when the protected attribute is not provided explicitly. REVISE implements 13 different metrics, some of which are very similar to what we described in the previous paragraphs. We are going to describe two of the most relevant: (i) scene diversity $H(S)$ where $S$ is the scene attribute computed applying a pre-trained scene recognition algorithm (Zhou et al., 2018); and (ii) appearance differences, computed by extracting features using a feature extractor of images with a same object/scene but a different protected attribute value and then training an SVM classifier on the extracted features to see whether it could learn different representations of subjects with different protected attribute values in the same context.

### 3.2. Biased image representations

The following methods analyse the distances and geometric relations among images exploiting their representation in a lower-dimensional space to discover the presence of bias.

*Distance-based.* Kärkkäinen and Joo (2021) proposed a simple measure of diversity for testing their face dataset. They studied the distribution of pairwise L1-distances calculated after embedding the images in a lower-dimensional space using a neural network pre-trained on a different benchmark face dataset. If such distribution is skewed towards high pairwise distances, it means the data show high diversity. Nonetheless, such an analysis is heavily influenced by the embedding. For instance, faces in a "white-oriented" dataset were also well separated in the embedding space, probably because the neural network used for the embedding had also been trained using a similarly biased dataset.

Steed and Caliskan (2021) developed a method for addressing human-like biases in the latent image representation of unsupervised generative models inspired by bias detection methods in Natural Language Processing (Caliskan et al., 2017). They discovered that the biases found in the latent space of two big models trained on ImageNet (Russakovsky et al., 2015) match with human-like biases. The authors measured bias by looking at associations between semantic concepts (for example, man-career and woman-family) by measuring the cosine similarity among vectors in the latent space computed by applying the models to controlled samples of images that resemble those visual concepts. More precisely, given a model $f$ that maps images into a vector space $\mathbb{R}^d$, two sets of images $J$ and $K$ (e.g., photos of men and women, respectively), and two sets $A$ and $B$ of images

representing the concepts we want to measure the association with (e.g., photos of people at work and photos of people in familiar settings), we can measure the association of $J$ with $A$ and $K$ with $B$ in the following way:

$$s(J, K, A, B) = \sum_{j \in J} s(j, A, B) - \sum_{k \in K} s(k, A, B) \qquad (4)$$

where

$$s(w, A, B) = mean_{a \in A} cos(f(w), f(a)) - mean_{b \in B} cos(f(w), f(b)) \qquad (5)$$

Alternatively, we can measure the size of the association via the Cohen's $d$:

$$d = \frac{mean_{j \in J} s(j, A, B) - mean_{k \in K} s(k, A, B)}{std_{w \in J \cup K} s(w, A, B)} \qquad (6)$$

The authors found out that the representations they analysed contained several biased associations that resemble human cognitive biases (e.g., the association between flower/insects with pleasant/unpleasant respectively, male/female with career/family or white/black with tools/weapons). Nevertheless, it is not clear how many of those associations were present in the original training data and to what degree this is the responsibility of the CV models used for computing the representations. While this method for measuring biased associations is model agnostic, in the sense that it can be applied to any possible model, its results heavily depend on the learned representations and the employed models.

*Other.* Balakrishnan et al. (2020) developed a method for assessing *algorithmic* biases in image classifiers via computing the causal relationship between the protected attribute and the output of a classifier. In doing so, they developed a method for intervening on the image attributes: they assumed to have a generator, such as a Generative Adversarial Network (Goodfellow et al., 2014), that generates images from latent vectors. They also assumed to have learned hyper-planes in the latent space that separate different attributes. Sampling points along the direction orthogonal to an attribute's hyper-plane gives a modified version of the original image with respect to that attribute. The authors noted that such interventions are not entirely disentangled: for example, adding long hair to the images of white males also adds the beard, and changing the gender attributes to the images of black males adds earrings. This is probably due to datasets bias which is then detected as a side effect of the transformation described above. Note that, since such transformation is the result of geometric operations, it means that the bias is encoded in the geometry of the latent space. It would be interesting to study to what extent these manipulations of the latent space can be used as a bias exploration tool.

### 3.3. Cross-dataset bias detection

Methods in this category derive from the realisation that the issue of generalisation in CV might be due to bias. Researchers in the field of object detection are usually able to tell with fair accuracy which famous benchmark dataset an image comes from Torralba and Efros (2011). This means that each dataset carries some sort of "signature" (bias) that makes the provenance of its images easily distinguishable and affects the ability of the models to generalise well. The methods described in the following aim to detect such signatures by comparing different datasets. Note that the works of Tommasi et al. (2015) and López-López et al. (2019) could also fit Section 3.2.

*Cross-dataset generalisation.* Torralba and Efros (2011) tested bias in object detection datasets by answering the following question: "How well does a typical object detector trained on one dataset generalise when tested on a representative set of other datasets, compared with its performance on the native test set?" The assumption here is that if the performance on the native test set is much higher it means the datasets exhibit some *bias* that is learned by the object detector $f_D$.

Hence, let us consider a dataset of interest $D$ and other $n$ benchmarks $\{B_i\}$. They compare performance by computing:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{AP_{B_i}(f_D)}{AP_D(f_D)} \tag{7}$$

The closer to 1 this score is, the more $f_D$ generalises well. Note that if this score is low, we can say that the $f_D$ does not generalise well and thus the dataset $D$ is probably biased, while if the score is close to 1 we can say that the datasets share a similar representation of the visual world. Furthermore, the authors also proposed a test, which they presented as a toy experiment but which has been utilised in many other studies, called *Name the dataset*. They trained a model to recognise the source dataset of a particular image: the greater the accuracy of this model, the more distinguishable and the more biased the datasets are. The methods described above have also been used by Tommasi et al. (2015), Panda et al. (2018) and Kärkkäinen and Joo (2021).

Tommasi et al. (2015) investigated the possibility of using CNN feature descriptors to address dataset bias. In particular, they replicated the experiments by Torralba and Efros (2011) and Khosla et al. (2012) (see next paragraph) using DeCAF features (Donahue et al., 2014). Furthermore, they slightly changed the measure used by Torralba and Efros for the evaluation of cross-generalisation:

$$CD = \sigma(\frac{1}{100}(AP_D(f_D) - \frac{1}{n}\sum_{i=1}^{n} AP_{B_i}(f_D))) \tag{8}$$

where $\sigma()$ is the sigmoid function.

*Other.* Khosla et al. (2012) proposed a method for both modelling and mitigating dataset bias. In particular, they train an SVM binary classifier on the union of a set of $n$ datasets $D_i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^{s_i}$, where $|D_i| = s_i$, $y_j^i \in \{-1, 1\}$ is a common class among all the $n$ datasets, and where $\mathbf{x}_j^i$ are feature vectors extracted via some feature extraction algorithms. The problem is framed as a multi-task classification (Evgeniou and Pontil, 2004) where the algorithm learns $n$ distinct hyper-planes $\mathbf{w}_i \cdot \mathbf{x}$ where $\mathbf{w}_i = \mathbf{w} + \Delta_i$. The vector $\mathbf{w}$, which is common to each dataset, models the common representation of the visual world while the specific biases of each of the $D_i$ datasets are modelled by the vectors $\Delta_i$. This is achieved via the following minimisation problem

$$\min_{\mathbf{w},\Delta_i} \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\left[\frac{\lambda}{2}\|\Delta_i\|^2 + \mathcal{L}(\mathbf{w}, \Delta_i)\right] \tag{9}$$

where

$$\mathcal{L}(\mathbf{w}, \Delta_i) = \sum_{j=1}^{s_i} C_1 \min(1, y_j^i \mathbf{w}\cdot \mathbf{x_j^i}) + C_2 \min(1, y_j^i(\mathbf{w}+\Delta_i)\cdot\mathbf{x_j^i}) \tag{10}$$

and $\lambda$, $C_1$ and $C_2$ are hyper-parameters. The authors proved that studying the vectors $\Delta_i$ gives useful semantic information about each dataset bias (e.g., they discovered that Caltech101 (Fei-Fei et al., 2004) has a strong preference on the side view of cars).

López-López et al. (2019) also investigated the use of feature descriptors to discover biases. In particular, they wanted to understand how the images in different datasets are distributed after embedding them in the same lower-dimensional space. Given $n$ datasets, $D_1, \ldots, D_n$ they sampled two sets of images for each of them $G_1, \ldots, G_n$ and $P_1, \ldots, P_n$, called respectively *gallery sets* and *probe sets*. Then, they computed the latent space applying a pre-trained feature descriptor $f$ to the gallery sets. After that, they computed the following probability:

$$\mathbb{P}(\mathbf{x}_* \in D_i | \mathbf{x} \in D_i) \tag{11}$$

where $\mathbf{x}$ is the feature vector of an image in the probe set $P_i$ and $\mathbf{x}_*$ is its nearest neighbour among the feature vectors of the images in the union of the gallery sets. If such probability is not $\frac{1}{n}$, it means that the nearest neighbours are not equally distributed among the $n$ datasets. Hence, there must be some *selection* bias.

### 3.4. Other methods

Other visual dataset bias discovery methods do not fit any of the three categories described above, and range from crowdsourcing frameworks to ad-hoc trained classification models.

*Model-based.* Model and Shamir (2015) proposed a simple method for addressing bias in object recognition datasets. They cropped a small central sub-image from each image in the original dataset. These cropped pictures were so small that humans could not recognise the objects in the pictures. Hence, if the attained performance of a model trained on such images was better than pure chance, the data contained distinguishable features spuriously correlated with the object categories. Then the data showed some kind of *selection* or *framing* (*capture*) bias.

Thomas and Kovashka (2019) studied the political framing bias of images by scraping images from online news outlets. Their idea was to train a semi-supervised tool for classifying images according to their political orientation. They labelled images according to the political orientation of the source. They also used information regarding the articles hosting the images by feeding the network with both the image and the document embedding of the article. Note that the proposed architecture incorporates textual information at training time but allows them to classify images without any additional information at testing time. Thus, this model can be used to understand if a specific image or a collection of images is biased towards a particular political faction. Moreover, the visual explanation of such a model could give semantic information on the political framing of the dataset.

Clark et al. (2020) proposed to use an ensemble classification algorithm for mitigating bias. The idea is to train a low-capacity network (i.e., with a low number of parameters) together with a high-capacity one (i.e., with more than double the parameters) so that the former learns spurious correlations while the latter learns to classify the data in an unbiased way. While this difference in capacity and the ensemble training encourages the two models to learn different patterns, there is still the possibility for the high-capacity model to learn simpler patterns and hence bias. To avoid this, the two networks are trained to result in conditionally independent outputs. This will be an incentive for the ensemble to isolate simple and complex patterns. While the authors use this method with the only purpose of mitigating algorithmic bias, studying the lower-capacity model can give information on spurious correlations in the training dataset, highlighting selection/framing bias.

Lopez-Paz et al. (2017) proposed a method to discover "causal signals" among objects appearing in image datasets. In particular, given a set of images $D = \{d_i\}$, they score the presence of certain objects/attributes $A$, $B$ by using an object detector or a feature extractor, respectively. Then they apply a Neural Causation Coefficient (NCC, (Lopez-Paz et al., 2015)) network, a neural network that takes a bag of samples $\{(a_i, b_i)_{i=1}^m\}$ as input and returns a score $s \in [0, 1]$ where $s = 0$ means that $A$ causes $B$ and $s = 1$ means instead that $B$ causes $A$. Note that, while not every causal relationship is a source of bias, some might be. Moreover, such causal relationships can be thought of as a by-product of the selection of the subject. Hence this method can detect selection bias.

*Human-based.* Hu et al. (2020) proposed a non-automated approach for bias discovery. Their method consists of a three-step crowdsourcing workflow for bias detection (selection and framing, according to our categorisation). In order to avoid the complexity of free text description, in the first step, workers are presented with a batch of images and asked to describe the similarities among the images of the batch via a question-answer pair (e.g., if every image of the batch shows only white airplanes, and then there is some selection bias, the worker would label the batch with the following question-answer pair: What colour are the airplanes in the images? White). In a second step, each worker is asked to answer some of the questions collected in the first phase based on different batches to confirm the presence of such biases. Finally, the

**Table 5**
Summary of the datasets described in Section 4.

| Paper | Year | Size | Protected attributes | Pre-existing Content | Labelling Process | Selection | Framing | Label |
|---|---|---|---|---|---|---|---|---|
| Buolamwini and Gebru (2018) | 2018 | 1.2K images | Binary gender; skin tone | Yes | Human annotators; experts; additional information | • | | • |
| Kärkkäinen and Joo (2021) | 2021 | 108.5K images | Age; binary gender; race | Yes | Crowd workers | • | | |
| Merler et al. (2019) | 2019 | 970K images | Age; Binary gender; skin tone | Yes | Machine annotators; crowd workers | • | • | |
| Georgopoulos et al. (2020) | 2020 | 41K images; 44K videos | Age; binary gender | Yes | Human annotators; machine annotators | • | • | |
| Barbu et al. (2019) | 2019 | 50K images | – | No | – | | • | |
| Wu et al. (2020) (Inclusive Benchmark) | 2020 | 12K images | Binary gender; race | Yes | Additional information | • | | |
| Wu et al. (2020) (Non-binary Gender Benchmark) | 2020 | 2K images | Non-binary gender; race | Yes | Additional information | • | | • |
| Hazirbas et al. (2021) | 2021 | 45.1K videos | Age; Non-binary gender; skin tone | No | Human annotators; self-provided labels | • | • | • |

workers are asked to evaluate whether the statements are true in the real visual world or, on the contrary, constitute some biases (e.g., is it true that every airplane is white? If the answer is yes, this is not considered an instance of bias). Note that since this last step is based on "common sense knowledge and subjective belief" (Hu et al., 2020), it heavily relies upon workers' backgrounds and biases.

### 3.5. Discussion

The reduction to tabular data is a reasonable and effective way of discovering bias. These methods could leverage the great amount of work that has been already done in the field of Fair Machine Learning for tabular data. Nevertheless, it seems that for what concerns visual data, the methods used are relatively simplistic. Most of the works just look for balance in the protected attribute or compare the distribution with respect to other features. Furthermore, these methods heavily rely on labels that are either attached to the data or automatically extracted. Hence, any labelling bias affects such discovery methods and should be taken into account.

In a complementary way, the image representation methods reduce the problem to bias detection in a lower-dimensional space instead of reducing it to tabular data. While this could better capture the complexity of visual content, such methods are necessarily influenced by both the models used to compute the representation and the metric used to compute distances in it. Moreover, they are usually harder to apply since they need some kind of supervision for computing the latent space.

Despite their historical importance, cross-dataset detection methods suffer from major issues. First, they are only applicable when several comparable datasets are available, which is not often practical. Second, they might help to unveil the presence of bias, but without further inspections, they cannot reveal what kind of bias it is. Note that, while these methods might be useful to get an idea of the existence of some biases in a visual dataset, they are of little or no use if we want to discover bias *within* the dataset, for example, if we want to understand whether there are discriminatory differences between men and women in the same dataset.

Regarding those methods that do not fit any of the above-mentioned categories, we cannot outline common pros and cons as they are very problem/domain-specific.

## 4. Bias-aware visual data collection

In the following, we are going to describe some attempts to construct datasets where the existence of bias was taken into account during the dataset creation process. These datasets were constructed for specific purposes and were probably not thought of as universally bias-free datasets. Nonetheless, analysing which biases have been removed

and which have been not might be helpful to understand the general challenge of bias in visual datasets. We summarise the content of this section in Table 5. Furthermore, we propose a checklist for helping the collection of bias-aware visual datasets ( Table 6). Note that, while the previous section systematically reviewed the literature on bias discovery and quantification in visual data, this section is meant as a case study. Therefore, it is more limited in its scope.

### 4.1. Datasets

Buolamwini and Gebru (2018) released the Pilot Parliaments Benchmark (PPB) dataset. PPB is a *face dataset* constructed by collecting the photos of members of six different national parliaments. The authors aimed to collect balanced data regarding the gender of the subjects and their skin colour. To do so, they selected three nations from African countries (Rwanda, Senegal, and South Africa) and three from European countries (Iceland, Finland, and Sweden) according to the gender parity rank.[11] among their Members of Parliament (MP). The data have been labelled by three annotators (including the authors) according to (binary) gender appearance and the Fitzpatrick skin type (ranging from I to IV, these labels are used by dermatologist as a gold standard for skin classification). The definitive skin labels were provided by a board-certified dermatologist, while the definitive gender labels were also determined based on the title, prefix, or the name of the parliamentarians (note that using names as a proxy of gender can cause label bias (Karimi et al., 2016)). While the data collection process described above resulted in a much more balanced dataset compared to other famous benchmarks – e.g., Adience (Eidinger et al., 2014); IJB-A (Klare et al., 2015) – it is still not free from possible biases. For example, the selection process targets a small number of African and northern European countries to ensure gender and skin tone balance. Nevertheless, it completely excludes, for instance, Asian and South American countries. Moreover, MPs are likely to be middle-aged people, which could also exclude young and older people from the selection. On the framing bias side, different countries might have different standards/dress codes for their MPs' official portraits, which could turn into some biases as well.

Kärkkäinen and Joo (2021) collected a *face dataset* emphasising, in particular, the balance in terms of age, gender, and race. They relied on a crowdsourcing workflow for annotating the images. In particular, they asked three different workers to label the images according to gender, age group, and race. They kept the label if there was a 2/3 accordance. Otherwise, they would have proposed the image to other three workers and discarded it if again it resulted in three different

---

[11] https://data.ipu.org/women-ranking?month=6&year=2021 Last visited 23.03.2022.

judgements. We can spot two sources of label and selection bias, respectively: first, we cannot be sure that the workers are able to determine the three labels homogeneously across every sub-group[12]; second, discarding the photos the workers cannot agree on might result in the missed selection of a certain group of people whose characteristics are difficult to determine for the workers. Finally, the taxonomy of races used by the authors (White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino) already introduces a form of label bias. While it is derived from the taxonomy commonly used by the US Census Bureau and might be descriptive of the composition of the US population, it hardly captures the complexity of human diversity.

Diversity in Face (DiF) (Merler et al., 2019) and KANFaces (Georgopoulos et al., 2020) are two *face datasets* that try to address the issue of bias by ensuring as much diversity as possible using the diversity measures proposed by Merler et al. (2019) that we reviewed in Section 3. The attributes they control the diversity for are: age, gender, skin tone, a set of craniofacial ratios, and pose. The authors also took into account a metric of illumination. By collecting diverse data, the authors try to avoid selection bias (in the case of age, gender, and skin tone) and some framing bias (in the case of pose and illumination).

Barbu et al. (2019) made an attempt to avoid framing bias in large-scale *object detection datasets*. In particular, they added controls for object rotations, viewpoints and background by asking crowd workers to photograph objects in their homes in a natural setting according to the instructions given by the authors. While this resulted in a much more diverse dataset (the authors used ImageNet (Russakovsky et al., 2015) as a comparison), because of the above controls, the objects appear only in indoor context, are rarely occluded, and are often centre-aligned. Thus, it seems that specific framing biases have been avoided, while the collection procedure has introduced others. Also, the authors removed some classes from the dataset for reasons that range from privacy concerns (e.g., "people") or because they were not easy to move and photograph in different settings (e.g., "beds"). In principle, this might generate some selection bias (more specifically, negative class bias) since the absence of those objects could make the negative classes less representative.

Wu et al. (2020) collected two benchmark datasets: the Inclusive Benchmark Database (IBD), and Non-binary Gender Benchmark Database (NGBD).[13] IBD contains 12,000 images of 168 different subjects, 21 of whom identify as LGBTQ. The geographic provenance of the subjects in the dataset is balanced. NGBD contains 2,000 images of 67 unique subjects. The subjects are public figures whose gender identity is known. Thus, the database contains multiple gender identities (namely: *non-binary*, *genderfluid*, *genderqueer*, *gender non-conforming*, *agender*, *gender neutral*, *gender-less*, *third gender*, and *queer*). The authors themselves identify two major risks of label bias: first, "Gender identity has its multifaceted aspects that a simple label could not categorize" (Wu et al., 2020) (the authors identify the problem of modelling gender as a continuum as a direction for future work); and, second, "Gender is a complex socio-cultural construct and an internal identity that is not necessarily tied to physical appearances" (Wu et al., 2020).

Hazirbas et al. (2021) proposed the Casual Conversations Dataset for evaluating the performance of CV models across different demographic categories. Their dataset is composed of 3,011 subjects and contains over 45,000 videos, with an average of 15 videos per person. The videos were recorded in multiple US states with a diverse set of adults in various age, gender and apparent skin tone groups. This work represents probably the greatest effort to build a balanced dataset addressing

both selection and framing bias (in the form of the illumination of videos). Nevertheless, some forms of imbalance remain. For example, most videos present bright lighting conditions; and most subjects are labelled as either male or female (with just the 0.1% of the participants that identify as "Others" and 2.1% whose gender is unknown). The label bias that the use of categories such as gender, age, and race could create is overcome by asking the participants their age and gender and using the Fitzpatrick Skin Type instead of race. Nonetheless, the authors declare that there are "videos in which two subjects are present simultaneously" but that they provide only one set of labels which might also be a form of label bias. Last, we note that the subjects are all from the US, which represents a serious selection bias as the US population hardly represents the whole of humankind.

*4.2. Discussion*

As highlighted by the cases analysed in the previous section, dealing with bias in visual data is not a trivial task. In particular, collecting bias-aware visual datasets might be incredibly challenging. Thus, we propose a checklist ( Table 6) to help scientists and practitioners spot and make explicit possible biases in the data they collect or use. Furthermore, we add a brief discussion on pre-processing mitigation strategies. Note that bias mitigation is an area on its own and would require an entire review to be adequately discussed. Therefore, we will just provide the interested reader with some hints and references.

Our checklist is inspired by previous work on documentation and reflexive data practices (Gebru et al., 2018; Miceli et al., 2021), but adds several questions specific to selection, framing, and label bias because they have their own peculiarities and must be analysed separately. We start with a general set of questions on the dataset purpose and the collection procedure. Then, we continue with questions specific to selection, framing, and label bias. In particular, we ask whether the selection of subjects generates any imbalance or lack of diversity, whether there are spurious correlations or harmful framing, whether fuzzy categories are used for labelling, and whether the labelling process contributes to inserting biases.

We want to stress that careful data collection and curation are probably the most effective mitigation strategies for all the three types of bias presented in Section 2; see, for example, the mitigation insights described in Wang et al. (2020). However, selection bias seems to be the easiest to mitigate using standard pre-processing techniques such as re-sampling or re-weighting. Note that several image datasets have a long-tail distribution of objects, and hence pre-processing mitigation techniques have to take that into account (Zhang et al., 2021), especially when many objects can co-occur in the same image. When the label bias is generated by poor operationalisation of unobservable theoretical construct (Jacobs and Wallach, 2021), the entire labelling must be reconsidered. Otherwise, when bias is caused by the presence of synonyms or incorrect labelling, an effective data cleaning could help the mitigation. When it appears in the form of capture bias, framing bias can be mitigated by either re-sampling techniques that ensure diversity (in pose and lighting, for example) or via data augmentation. When the framing bias relates to the messages carried by the images, automatic pre-processing strategies are probably less effective as the way an image is interpreted and the message it conveys are not universal, which gives rise to the need for careful human inspection and analysis.

## 5. Conclusions and research outlook

The aim of this survey was threefold: first, to provide a description of different types of bias and to illustrate the processes through which they affect CV applications and visual datasets; second, to perform a systematic review of methods for bias discovery in visual datasets; and, third, to describe existing attempts to collect visual datasets in a bias-aware manner. We showed how the problem of bias is pervasive in CV;

---

[12] see the interesting Twitter thread from Sacha Costanza-Chock (@schock), https://twitter.com/schock/status/1346478831255789569 (Last visited 23.03.2022) for an on-point discussion about this issue.

[13] While collecting inclusive data is a noble purpose, the use the authors made of their dataset (extending gender classification algorithms to non-binary genders) should be strongly discouraged as it might be used to target already marginalised groups of people.

**Table 6**
Checklist for bias-aware visual data collection.

|  | Description |
|---|---|
| General | What are the purposes the data are collected for? (Gebru et al., 2018) |
|  | Are there uses of the data that should be discouraged because of possible biases? (Gebru et al., 2018) |
|  | Does the dataset's licence agreement explicitly restrict or forbid those uses? |
|  | What kind of bias can be inserted by the way the collection process is designed? (Gebru et al., 2018) |
| Selection bias | Do we need balanced data or statistically representative data? |
|  | Are the negative sets representative enough? |
|  | Is the dataset representative enough? |
|  | Is there any group of subjects that is systematically excluded from the data? |
|  | Do the data come from or depict a specific geographical area (e.g., mainly western countries)? |
|  | Does the selection of the subjects create any spurious associations? |
|  | Will the dataset remain representative for its foreseeable lifetime? |
| Framing bias | Are there any spurious correlation that can contribute to framing different subjects in different ways? |
|  | Is there any biases due to the way images/videos are captured? |
|  | Did the capture induce some behaviour in the subjects (e.g. smiling when photographed)? |
|  | Are there any images that can possibly convey different messages depending on the viewer? (e.g., specific cultural references) |
|  | Are subjects of a certain group depicted in a particular context more often than others? (e.g., female subjects in household settings) |
|  | Do the data agree with harmful stereotypes? |
| Label bias | If the labelling process relies on machines: have their biases been taken into account? (e.g., biases of pre-trained classification models) |
|  | If the labelling process relies on human annotators: is there an adequate and diverse pool of annotators? Have their possible biases been taken into account? |
|  | If the labelling process relies on crowd sourcing: are there any biases due to the workers' access to crowd sourcing platforms? |
|  | Do we use fuzzy labels? (e.g., race or gender) |
|  | Do we operationalise any unobservable theoretical constructs/use proxy variables? (Jacobs and Wallach, 2021) |

It accompanies the whole life cycle of visual content, involves several actors, and re-enters the life cycle through biased CV algorithms. One of our major contributions has been to provide a detailed description of the different manifestations (selection, framing, and label) of bias in visual data, along with several examples. We also went further by providing a sub-categorisation ( Table 1) which also includes several categories of bias that are commonly described in Statistics, Health studies, or Psychology adapting them to the context of visual data.

The systematic review in Section 3 allowed us to draw some consideration on the state of the art in bias discovery methods for visual data and to outline some possible future streams of research. The vast majority of the methods we reviewed used the strategy of reducing the problem to tabular data (Section 3.1). While it seems a natural option as it allows leveraging the wealth of techniques from the Fair Accountable and Transparent Machine Learning (FATML) literature, using different representations of visual data would certainly open new unexplored research directions. For example, representing an image as a scene graph (Johnson et al., 2015), would allow the use of bias detection techniques used for rankings and graphical data (Krasanakis et al., 2020; Pitoura et al., 2021). Moreover, since scene graphs can be thought of as small-scale knowledge graphs, this representation would allow both the integration of additional knowledge (e.g., from Wikipedia) and the use of methods for bias detection in knowledge graph embeddings (Bourli and Pitoura, 2020).

Balakrishnan et al. (2020) showed that a latent representation of data encapsulates bias in its geometry. This was also noted by Bolukbasi et al. (2016) in their work on bias in word embeddings. Nonetheless, the study of this relationship between the geometry of latent spaces and bias is usually limited to variations of cosine similarity. Hence, a promising research direction would be to exploit more sophisticated tools for studying such relationships, e.g., Topological Data Analysis (Chazal and Michel, 2021).

Some of the most influential works on bias detection in images, e.g., Torralba and Efros (2011), focused on object detection datasets.

In such datasets, the long-tail distribution of objects is very common and evaluating bias in these datasets is difficult because the long tail distracts from the focal points. Moreover, the co-occurrence of different objects has obvious implications for the effectiveness of mitigation methods (Zhang et al., 2021) for imbalanced data (e.g., oversampling, under-sampling), and therefore it should be studied more deeply.

Most of the reviewed papers focus on images. Only few of them study bias in videos. This opens the road to many possibilities (for example, applying methods from time-series and multimodal analysis to the bias detection domain). Furthermore, we noted that selection and framing bias are the two types of bias that are more commonly taken into account. Nevertheless, label bias is often very concerning (Section 2.3) and hence more research is needed to fill this gap.

Similarly, for obvious simplicity reasons, most related works concentrate on a single protected attribute (an exception to this is the work of Buolamwini and Gebru (2018)). Hence, multi-attribute fairness studies would enrich the literature in this respect. We also noted that most of the works we reviewed in Section 4 focus on facial data. This is probably due to the concerning applications of face detection/recognition algorithms. Therefore, there is room for improving datasets and establishing data collection practices in other fields, including medical imaging, self-driving cars, etc. Nonetheless, we would like to stress that no amount of debiasing can save some applications from their inherent flaws (e.g., criminality-from-face detection as described in Section 2.3).

We finally reviewed several attempts to collect bias-aware data in Section 4 and concluded that there is no such thing as bias-free data. Hence, it is of utmost importance, along with the development of reliable bias discovery tools, that researchers and practitioners become aware of the biases of the datasets they collect and make them explicit in a standardised way (Gebru et al., 2018; Miceli et al., 2021). In Table 6, we outline a checklist for the collection of visual data. We believe that having such a guide will help practitioners and scientists spot possible causes of bias, collect data that is as less biased as possible, and be aware of such biases during their analysis.

## Declaration of competing interest

## Acknowledgements

## References

Angwin, Julia, Larson, Jeff, Mattu, Surya, Kirchner, Lauren, 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Archer, Dane, Iritani, Bonita, Kimes, Debra D., Barrios, Micheal, 1983. Face-ism: Five studies of sex differences in facial prominence. J. Personal. Soc. Psychol. 45, 725–735.

Balakrishnan, Guha, Xiong, Yuanjun, Xia, Wei, Perona, Pietro, 2020. Towards causal benchmarking of bias in face analysis algorithms. In: Vedaldi, Andrea, Bischof, Horst, Brox, Thomas, Frahm, Jan-Michael (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII. In: Lecture Notes in Computer Science, vol. 12363, Springer, pp. 547–563. http://dx.doi.org/10.1007/978-3-030-58523-5_32.

Bandy, Jack, 2021. Problematic machine behavior: A systematic literature review of algorithm audits. Proc. ACM Hum.-Comput. Interact. 5 (CSCW1), http://dx.doi.org/10.1145/3449148.

Baraniuk, Richard, Donoho, David L., Gavish, Matan, 2020. The science of deep learning. Proc. Natl. Acad. Sci. 117, 30029–30032.

Barbu, Andrei, Mayo, David, Alverio, Julian, Luo, William, Wang, Christopher, Gutfreund, Dan, Tenenbaum, Josh, Katz, Boris, 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In: Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F., Fox, E., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., URL https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf.

Barbujani, Guido, Colonna, Vincenza, 2011. Genetic basis of human biodiversity: An update. Springer, pp. 97–119. http://dx.doi.org/10.1007/978-3-642-20992-5_6.

Bartlett, Robert, Morse, Adair, Stanton, Richard, Wallace, Nancy, 2019. Consumer-Lending Discrimination in the FinTech Era. Working Paper 25943, In: Working Paper Series, National Bureau of Economic Research, http://dx.doi.org/10.3386/w25943, URL http://www.nber.org/papers/w25943.

Berendt, Bettina, Gandon, Fabien, Halford, Susan, Hall, Wendy, Hendler, Jim, Kinder-Kurlanda, Katharina E., Ntoutsi, Eirini, Staab, Steffen, 2021. Web futures: Inclusive, intelligent, sustainable the 2020 manifesto for web science (dagstuhl perspectives workshop 18262). Dagstuhl Manifestos 9 (1), 1–42. http://dx.doi.org/10.4230/DagMan.9.1.1.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James, Saligrama, Venkatesh, Kalai, Adam, 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. NIPS '16, Curran Associates Inc., Red Hook, NY, USA, pp. 4356–4364.

Bourli, Styliani, Pitoura, Evaggelia, 2020. Bias in knowledge graph embeddings. In: Atzmüller, Martin, Coscia, Michele, Missaoui, Rokia (Eds.), IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2020, the Hague, Netherlands, December 7-10, 2020. IEEE, pp. 6–10. http://dx.doi.org/10.1109/ASONAM49781.2020.9381459.

Bowyer, Kevin W., King, Micheal C., Scheirer, Walter J., Vangara, Kushal, 2020. The "criminality from face" illusion. IEEE Trans. Technol. Soc. 1 (4), 175–183. http://dx.doi.org/10.1109/TTS.2020.3032321.

Buolamwini, Joy, Gebru, Timnit, 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, Sorelle A., Wilson, Christo (Eds.), Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. In: Proceedings of Machine Learning Research, vol. 81, PMLR, pp. 77–91. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Caliskan, Aylin, Bryson, Joanna J., Narayanan, Arvind, 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356 (6334), 183–186. http://dx.doi.org/10.1126/science.aal4230, URL https://science.sciencemag.org/content/356/6334/183 Publisher: American Association for the Advancement of Science _eprint: https://science.sciencemag.org/content/356/6334/183.full.pdf.

Chazal, Frédéric, Michel, Bertrand, 2021. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. Front. Artif. Intell. 4, http://dx.doi.org/10.3389/frai.2021.667963, URL https://www.frontiersin.org/article/10.3389/frai.2021.667963.

Chouldechova, Alexandra, 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data 5 2, 153–163.

Clark, Christopher, Yatskar, Mark, Zettlemoyer, Luke, 2020. Learning to model and ignore dataset bias with mixed capacity ensembles. In: Cohn, Trevor, He, Yulan, Liu, Yang (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020. Association for Computational Linguistics, pp. 3031–3045. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.272.

Coleman, Renita, 2010. Framing the pictures in our heads: Exploring the framing and agenda-setting effects of visual images. Doing Frame Anal. Empirical Theor. Perspect. 233–261.

Cook, Cynthia M., Howard, John J., Sirotin, Yevgeniy B., Tipton, Jerry L., Vemury, Arun R., 2019. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. IEEE Trans. Biometrics Behav. Identity Sci. 1 (1), 32–41. http://dx.doi.org/10.1109/TBIOM.2019.2897801.

Corradi, Laura, 2012. Specchio Delle Sue Brame: Analisi Socio-Politica Delle Pubblicità : Genere, Classe, Razza, Età Ed Eterosessismo. In: Sessismo e razzismo. Saggi, Ediesse, URL https://books.google.gr/books?id=ckGfpwAACAAJ.

Cramer, Maria, Levenson, Micheal, 2021. Yearbook photos of girls were altered to hide their chests. The NY Times URL https://www.nytimes.com/2021/05/23/us/yearbook-photos-st-johns-girls-altering.html.

Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, Fei-Fei, Li, 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Donahue, Jeff, Jia, Yangqing, Vinyals, Oriol, Hoffman, Judy, Zhang, Ning, Tzeng, Eric, Darrell, Trevor, 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In: Xing, Eric P., Jebara, Tony (Eds.), Proceedings of the 31st International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 32, PMLR, Bejing, China, pp. 647–655, URL http://proceedings.mlr.press/v32/donahue14.html.

Drozdowski, Pawel, Rathgeb, Christian, Dantcheva, Antitza, Damer, Naser, Busch, Christoph, 2020. Demographic bias in biometrics: A survey on an emerging challenge. IEEE Trans. Technol. Soc. 1 (2), 89–103. http://dx.doi.org/10.1109/TTS.2020.2992344.

Dulhanty, Chris, Wong, Alexander, 2019. Auditing ImageNet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. arXiv arXiv:1905.01347.

Eidinger, Eran, Enbar, Roee, Hassner, Tal, 2014. Age and gender estimation of unfiltered faces. IEEE Trans. Inf. Forensics Secur. 9, 2170–2179. http://dx.doi.org/10.1109/TIFS.2014.2359646.

Entman, Robert M., 1993. Framing: Toward clarification of a fractured paradigm. J. Commun. 43 (4), 51–58. http://dx.doi.org/10.1111/j.1460-2466.1993.tb01304.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.1993.tb01304.x.

Evgeniou, Theodoros, Pontil, Massimiliano, 2004. Regularized multi–task learning. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '04, Association for Computing Machinery, New York, NY, USA, pp. 109–117. http://dx.doi.org/10.1145/1014052.1014067.

Fei-Fei, Li, Fergus, Rob, Perona, Pietro, 2004. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop. pp. 178–178. http://dx.doi.org/10.1109/CVPR.2004.383.

Gebru, Timnit, Morgenstern, Jamie H., Vecchione, Briana, Vaughan, Jennifer Wortman, Wallach, H., Daumé, Hal, Crawford, Kate, 2018. Datasheets for datasets. arXiv arXiv:1803.09010.

Georgopoulos, Markos, Panagakis, Yannis, Pantic, Maja, 2020. Investigating bias in deep face analysis: The KANFace dataset and empirical study. Image Vis. Comput. 102, 103954. http://dx.doi.org/10.1016/j.imavis.2020.103954, URL https://www.sciencedirect.com/science/article/pii/S026288562030086X.

Goldenfein, Jake, 2019. The profiling potential of computer vision and the challenge of computational empiricism. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. In: FAT* '19, Association for Computing Machinery, New York, NY, USA, pp. 110–119. http://dx.doi.org/10.1145/3287560.3287568.

Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, Bengio, Yoshua, 2014. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. NIPS '14, MIT Press, Cambridge, MA, USA, pp. 2672–2680.

Hanna, Alex, Denton, Emily, Smart, Andrew, Smith-Loud, Jamila, 2020. Towards a critical race methodology in algorithmic fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, New York, NY, USA, pp. 501–512. http://dx.doi.org/10.1145/3351095.3372826.

Hazirbas, Caner, Bitton, Joanna, Dolhansky, Brian, Pan, Jacqueline, Gordo, Albert, Canton-Ferrer, Cristian, 2021. Towards measuring fairness in AI: the casual conversations dataset.

Hernán, Miguel A., Robins, Jamie M., 2020. Causal Inference: What if. Chapman & Hall/CRC.

Heuer, Chealse A., McClure, Kimberly J., Puhl, Rebecca M., 2011. Obesity stigma in online news: a visual content analysis. J. Health Commun. 16 (9), 976–987.

Hill, Kashmir, 2020. Wrongfully accused by an algorithm. The NY Times URL https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hu, Xiao, Wang, Haobo, Vegesana, Anirudh, Dube, Somesh, Yu, Kaiwen, Kao, Gore, Chen, Shuo-Han, Lu, Yung-Hsiang, Thiruvathukal, George K., Yin, Ming, 2020. Crowdsourcing detection of sampling biases in image datasets. In: Proceedings of the Web Conference 2020. WWW '20, Association for Computing Machinery, New York, NY, USA, pp. 2955–2961. http://dx.doi.org/10.1145/3366423.3380063.

Huang, Gary B., Mattar, Marwan, Berg, Tamara, Learned-Miller, Eric, 2008. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France, URL https://hal.inria.fr/inria-00321923.

Jacobs, Abigail Z., Wallach, Hanna, 2021. Measurement and fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery, New York, NY, USA, pp. 375–385. http://dx.doi.org/10.1145/3442188.3445901.

Jamil, Jameela, 2018. Viewpoint: Jameela jamil on why airbrushing should be illegal. BBC URL https://www.bbc.com/news/world-46349307.

Jang, Ji Yoon, Lee, Sangyoon, Lee, Byungjoo, 2019. Quantification of gender representation bias in commercial films based on image analysis. Proc. ACM Hum.-Comput. Interact. 3 (CSCW), http://dx.doi.org/10.1145/3359300.

Jiang, Heinrich, Nachum, Ofir, 2020. Identifying and correcting label bias in machine learning. In: AISTATS.

Johnson, Justin, Krishn, Ranjay, Stark, Micheal, Li, Li-Jia, Shamma, Davod A., Bernstein, Micheal S., Fei-Fei, Li, 2015. Image retrieval using scene graphs. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3668–3678. http://dx.doi.org/10.1109/CVPR.2015.7298990.

Karimi, Fariba, Wagner, Claudia, Lemmerich, Florian, Jadidi, Mohsen, Strohmaier, Markus, 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In: Proceedings of the 25th International Conference Companion on World Wide Web. In: WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, pp. 53–54. http://dx.doi.org/10.1145/2872518.2889385.

Kärkkäinen, Kimmo, Joo, Jungseock, 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, pp. 1548–1558.

Kasirzadeh, Atoosa, Smart, Andrew, 2021. The use and misuse of counterfactuals in ethical machine learning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery, New York, NY, USA, pp. 228–236. http://dx.doi.org/10.1145/3442188.3445886.

Kay, Matthew, Matuszek, Cynthia, Munson, Sean A., 2015. Unequal representation and gender stereotypes in image search results for occupations. In: Begole, Bo, Kim, Jinwoo, Inkpen, Kori, Woo, Woontack (Eds.), Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015. ACM, pp. 3819–3828. http://dx.doi.org/10.1145/2702123.2702520.

Khosla, Aditya, Zhou, Tinghui, Malisiewicz, Tomasz, Efros, Alexei A., Torralba, Antonio, 2012. Undoing the damage of dataset bias. In: Proceedings of the 12th European Conference on Computer Vision - Volume Part I. ECCV '12, Springer-Verlag, Berlin, Heidelberg, pp. 158–171. http://dx.doi.org/10.1007/978-3-642-33718-5_12.

Kim, Byungju, Kim, Hyunwoo, Kim, Kyungsu, Kim, Sungjin, Kim, Junmo, 2019. Learning not to learn: Training deep neural networks with biased data. In: The IEEE Conference on Computer Vision and Pattern Recognition. CVPR.

Kitchenham, Barbara, 2004. Procedures for performing systematic reviews. Keele, UK, Keele Univ. 33 (2004), 1–26.

Klare, Brendan, Burge, Mark James, Klontz, Joshua C., Bruegge, Richard W. Vorder, Jain, Anil K., 2012. Face recognition performance: Role of demographic information. IEEE Trans. Inf. Forensics Secur. 7 (6), 1789–1801. http://dx.doi.org/10.1109/TIFS.2012.2214212.

Klare, Brendan F., Klein, Ben, Taborsky, Emma, Blanton, Austin, Cheney, Jordan, Allen, Kristen, Grother, Patrick, Mah, Alan, Burge, Mark, Jain, Anil J., 2015. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1931–1939. http://dx.doi.org/10.1109/CVPR.2015.7298803.

Kleinberg, Jon M., Mullainathan, Sendhil, Raghavan, Manish, 2017. Inherent trade-offs in the fair determination of risk scores. In: Papadimitriou, Christos H. (Ed.), 8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA. In: LIPIcs, vol. 67, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 43:1–43:23. http://dx.doi.org/10.4230/LIPIcs.ITCS.2017.43.

Kofod-Petersen, Anders, 2012. How to do a structured literature review in computer science. Ver. 0.1. October 1.

Krasanakis, Emmanouil, Papadopoulos, Symeon, Kompatsiaris, Ioannis, 2020. Applying fairness constraints on graph node ranks under personalization bias. In: International Conference on Complex Networks and their Applications. Springer, pp. 610–622.

Krasin, Ivan, Duerig, Tom, Alldrin, Neil, Veit, Andreas, Abu-El-Haija, Sami, Belongie, Serge, Cai, David, Feng, Zheyun, Ferrari, Vittorio, Gomes, Victor, Gupta, Abhinav, Narayanan, Dhyanesh, Sun, Chen, Chechik, Gal, Murphy, Kevin, 2016. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. Dataset Available from https://Github.Com/Openimages.

Krug, Matthias, Niggemeier, Stefan, 2013. Exploring the boundaries of photo editing. Der Spiegel International URL https://www.spiegel.de/international/world/growing-concern-that-news-photos-are-being-excessively-manipulated-a-898509.html.

Kumar, Neeraj, Berg, Alexander C., Belhumeur, Peter N., Nayar, Shree K., 2009. Attribute and simile classifiers for face verification. In: IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009. IEEE Computer Society, pp. 365–372. http://dx.doi.org/10.1109/ICCV.2009.5459250.

LeCun, Yann, Bengio, Yoshua, Hinton, Geoffrey E., 2015. Deep learning. Nat. 521 (7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Liang, Lingyu, Lin, Luojun, Jin, Lianwen, Xie, Duorui, Li, Men, 2018. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In: 2018 24th International Conference on Pattern Recognition. ICPR, pp. 1598–1603. http://dx.doi.org/10.1109/ICPR.2018.8546038.

Lin, Tsung-Yi, Maire, M., Belongie, Serge J., Hays, James, Perona, P., Ramanan, D., Dollár, Piotr, Zitnick, C. L., 2014. Microsoft COCO: Common objects in context. In: ECCV.

López-López, Eric, Pardo, Xosé M., Regueiro, Carlos V., Iglesias, Roberto, Casado, Fernando E., 2019. Dataset bias exposed in face verification. IET Biometrics 8 (4), 249–258. http://dx.doi.org/10.1049/iet-bmt.2018.5224, arXiv:https://ietresearch.onlinelibrary.wiley.com/doi/pdf/10.1049/iet-bmt.2018.5224.

Lopez-Paz, David, Muandet, Krikamol, Schölkopf, Bernhard, Tolstikhin, Ilya O., 2015. Towards a learning theory of cause-effect inference. In: Bach, Francis R., Blei, David M. (Eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015. In: JMLR Workshop and Conference Proceedings, vol. 37, JMLR.org, pp. 1452–1461, URL http://proceedings.mlr.press/v37/lopez-paz15.html.

Lopez-Paz, David, Nishihara, Robert, Chintala, Soumith, Schölkopf, Bernhard, Bottou, Léon, 2017. Discovering causal signals in images. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 58–66. http://dx.doi.org/10.1109/CVPR.2017.14.

Merler, Michele, Ratha, Nalini, Feris, Rogerio S., Smith, John R., 2019. Diversity in faces. arXiv:1901.10436,

Merli, Roberto, Preziosi, Michele, Acampora, Alessia, 2018. How do scholars approach the circular economy? A systematic literature review. J. Cleaner Prod. 178, 703–722. http://dx.doi.org/10.1016/j.jclepro.2017.12.112, URL http://www.sciencedirect.com/science/article/pii/S0959652617330718.

Miceli, Milagros, Schuessler, Martin, Yang, Tianling, 2020. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. Proc. ACM Hum.-Comput. Interact. 4 (CSCW2), http://dx.doi.org/10.1145/3415186.

Miceli, Milagros, Yang, Tianling, Naudts, Laurens, Schuessler, Martin, Serbanescu, Dian, Hanna, Alex, 2021. Documenting computer vision datasets: An invitation to reflexive data practices. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21, Association for Computing Machinery, New York, NY, USA, pp. 161–172. http://dx.doi.org/10.1145/3442188.3445880.

Mirsky, Yisroel, Lee, Wenke, 2021. The creation and detection of deepfakes: A survey. ACM Comput. Surv. 54 (1), http://dx.doi.org/10.1145/3425780.

Model, Ian, Shamir, Lior, 2015. Comparison of data set bias in object recognition benchmarks. IEEE Access 3, 1953–1962. http://dx.doi.org/10.1109/ACCESS.2015.2491921.

Ntoutsi, Eirini, Fafalios, Pavlos, Gadiraju, Ujwal, Iosifidis, Vasileios, Nejdl, Wolfgang, Vidal, Maria-Esther, Ruggieri, Salvatore, Turini, Franco, Papadopoulos, Symeon, Krasanakis, Emmanouil, Kompatsiaris, Ioannis, Kinder-Kurlanda, Katharina, Wagner, Claudia, Karimi, Fariba, Fernández, Miriam, Alani, Harith, Berendt, Bettina, Kruegel, Tina, Heinze, Christian, Broelemann, Klaus, Kasneci, Gjergji, Tiropanis, Thanassis, Staab, Steffen, 2020. Bias in data-driven artificial intelligence systems - An introductory survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 10 (3), http://dx.doi.org/10.1002/widm.1356.

Panda, Rameswar, Zhang, Jianming, Li, Haoxiang, Lee, Joon-Young, Lu, Xin, Roy-Chowdhury, Ammit K., 2018. Contemplating visual emotions: Understanding and overcoming dataset bias. In: ECCV.

Paquette, Danielle, 2015. The uncomfortable truth about how we view working women, in one simple google search. The Washington Post URL https://www.washingtonpost.com/news/wonk/wp/2015/04/14/what-one-simple-google-search-tells-us-about-how-we-view-working-women/.

Peng, Yilang, 2018. Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. J. Commun. 68, http://dx.doi.org/10.1093/joc/jqy041.

Pitoura, Evaggelia, Stefanidis, Kostas, Koutrika, Georgia, 2021. Fairness in rankings and recommenders: Models, methods and research directions. In: 37th IEEE International Conference on Data Engineering, ICDE 2021, Chania, Greece, April 19-22, 2021. IEEE, pp. 2358–2361. http://dx.doi.org/10.1109/ICDE51399.2021.00265.

Prabhu, Vinay Uday, Birhane, Abeba, 2020. Large image datasets: A pyrrhic win for computer vision? arXiv arXiv:2006.16923.

Robinson, Joseph P., Livitz, Gennady, Henon, Yann, Qin, Can, Fu, Yun, Timoner, Samson, 2020. Face recognition: Too bias, or not too bias? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Micheal, Berg, Alexander C., Fei-Fei, Li, 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

Salakhutdinov, Ruslan, Torralba, Antonio, Tenenbaum, Joshua B., 2011. Learning to share visual appearance for multiclass object detection. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011. IEEE Computer Society, pp. 1481–1488. http://dx.doi.org/10.1109/CVPR.2011.5995720.

Shankar, Shreya, Halpern, Yoni, Breck, Eric, Atwood, James, Wilson, Jimbo, Sculley, D., 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In: NIPS 2017 Workshop: Machine Learning for the Developing World.

Singh, Krishna Kumar, Mahajan, Dhruv, Grauman, Kristen, Lee, Yong Jae, Feiszli, Matt, Ghadiyaram, Deepti, 2020. Don't judge an object by its context: Learning to overcome contextual bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR.

Steed, Ryan, Caliskan, Aylin, 2021. Image representations learned with unsupervised pre-training contain human-like biases. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, pp. 701–713.

Sweeney, Latanya, 2013. Discrimination in online ad delivery. Commun. ACM 56 (5), 44–54. http://dx.doi.org/10.1145/2447976.2447990.

Terhörst, Philipp, Fährmann, Daniel, Kolf, Jan Niklas, Damer, Naser, Kirchbuchner, Florian, Kuijper, Arjan, 2021a. MAAD-face: A massively annotated attribute dataset for face images. arXiv arXiv:2012.01030.

Terhörst, Philipp, Kolf, Jan Niklas, Huber, Marco, Kirchbuchner, Florian, Damer, Naser, Morales, Aythami, Fiérrez, Julian, Kuijper, Arjan, 2021b. A comprehensive study on face recognition biases beyond demographics. arXiv arXiv:2103.01592.

Thomas, Christopher, Kovashka, Adriana, 2019. Predicting the politics of an image using webly supervised data. In: NeurIPS.

Tommasi, Tatiana, Patricia, Novi, Caputo, Barbara, Tuytelaars, Tinne, 2015. A deeper look at dataset bias. In: Gall, Juergen, Gehler, Peter V., Leibe, Bastian (Eds.), Pattern Recognition - 37th German Conference, GCPR 2015, Aachen, Germany, October 7-10, 2015, Proceedings. In: Lecture Notes in Computer Science, vol. 9358, Springer, pp. 504–516. http://dx.doi.org/10.1007/978-3-319-24947-6_42.

Torralba, Antonio, Efros, Alexei A., 2011. Unbiased look at dataset bias. In: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011. IEEE Computer Society, pp. 1521–1528. http://dx.doi.org/10.1109/CVPR.2011.5995347.

Verma, Sahil, Rubin, Julia, 2018. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. FairWare '18, Association for Computing Machinery, New York, NY, USA, pp. 1–7. http://dx.doi.org/10.1145/3194770.3194776.

Wachinger, Christian, Rieckmann, Anna, Pölsterl, Sebastian, 2021. Detect and correct bias in multi-site neuroimaging datasets. Med. Image Anal. 67, 101879. http://dx.doi.org/10.1016/j.media.2020.101879, URL https://www.sciencedirect.com/science/article/pii/S1361841520302437.

Wang, Angelina, Narayanan, Arvind, Russakovsky, Olga, 2020. REVISE: A tool for measuring and mitigating bias in visual datasets. Euro. Conf. Comput. Vis. (ECCV).

Wang, Tianlu, Zhao, Jieyu, Yatskar, Mark, wei Chang, Kai, Ordonez, Vincente, 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 5309–5318. http://dx.doi.org/10.1109/ICCV.2019.00541.

Wilson, Benjamin, Hoffman, Judy, Morgenstern, Jamie, 2019. Predictive inequity in object detection. CoRR arXiv:1902.11097.

Wu, Wenying, Protopapas, Pavlos, Yang, Xheng, Michalatos, Panagiotis, 2020. Gender classification and bias mitigation in facial images. In: 12th ACM Conference on Web Science. WebSci '20, Association for Computing Machinery, New York, NY, USA, pp. 106–114. http://dx.doi.org/10.1145/3394231.3397900.

Yang, Kaiyu, Qinami, Klint, Fei-Fei, Li, Deng, Jia, Russakovsky, Olga, 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. In: FAT* '20, Association for Computing Machinery, New York, NY, USA, pp. 547–558. http://dx.doi.org/10.1145/3351095.3375709.

You, Quanzeng, Luo, Jiebo, Jin, Hailin, Yang, Jianchao, 2016. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. AAAI '16, AAAI Press, pp. 308–314.

Yu, Fisher, Chen, Haofeng, Wang, Xin, Xian, Wenqi, Chen, Yingying, Liu, Fangchen, Madhavan, Vashisht, Darrell, Trevor, 2020. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2633–2642. http://dx.doi.org/10.1109/CVPR42600.2020.00271.

Zhang, Yongshun, Wei, Xiu-Shen, Zhou, Boyan, Wu, Jianxin, 2021. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, the Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. AAAI Press, pp. 3447–3455, URL https://ojs.aaai.org/index.php/AAAI/article/view/16458.

Zhao, Jieyu, Wang, Tianlu, Yatskar, Mark, Ordonez, Vincente, Chang, Kai-Wei, 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp. 2979–2989. http://dx.doi.org/10.18653/v1/D17-1323, URL https://www.aclweb.org/anthology/D17-1323.

Zhou, Bolei, Lapedriza, Agata, Khosla, Aditya, Oliva, Aude, Torralba, Antonio, 2018. Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40 (6), 1452–1464. http://dx.doi.org/10.1109/TPAMI.2017.2723009.

Zhu, Xiangxin, Anguelov, Dragomir, Ramanan, Deva, 2014. Capturing long-tail distributions of object subcategories. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. IEEE Computer Society, pp. 915–922. http://dx.doi.org/10.1109/CVPR.2014.122.