Data Mining and Knowledge Discovery (2024) 38:281–312
https://doi.org/10.1007/s10618-023-00972-2

Studying bias in visual features through the lens of optimal transport

Simone Fabbrizzi[1,2] · **Xuan Zhao**[3,4] · **Emmanouil Krasanakis**[1] · **Symeon Papadopoulos**[1] · **Eirini Ntoutsi**[5]

Received: 10 February 2023 / Accepted: 27 July 2023 / Published online: 2 September 2023

Abstract

Computer vision systems are employed in a variety of high-impact applications. However, making them trustworthy requires methods for the detection of potential biases in their training data, before models learn to harm already disadvantaged groups in downstream applications. Image data are typically represented via extracted features, which can be hand-crafted or pre-trained neural network embeddings. In this work, we introduce a framework for bias discovery given such features that is based on optimal transport theory; it uses the (quadratic) Wasserstein distance to quantify disparity between the feature distributions of two demographic groups (e.g., women vs men). In this context, we show that the Kantorovich potentials of the images, which are a byproduct of computing the Wasserstein distance and act as "transportation prices", can serve as bias scores by indicating which images might exhibit distinct biased characteristics. We thus introduce a visual dataset exploration pipeline that helps auditors identify common characteristics across high- or low-scored images as potential sources of bias. We conduct a case study to identify prospective gender biases and demonstrate theoretically-derived properties with experiments on the CelebA and Biased MNIST datasets.

Keywords Fairness and bias · Computer vision · Optimal transport · Dataset exploration · Tools and frameworks · Wasserstein distance

Responsible editor: Charalampos Tsourakakis.

✉ Simone Fabbrizzi
simone.fabbrizzi@iti.gr

[1] CERTH-ITI, Thessaloniki, Greece

[2] Leibniz Universität, Hannover, Germany

[3] SCHUFA, Wiesbaden, Germany

[4] University of Tuebingen, Tuebingen, Germany

[5] Research Institute CODE, Bundeswehr University, Munich, Germany

1 Introduction

Computer Vision (CV) systems have widespread applications, but their performance depends on the quality of the training data. When these data exhibit biases, trained systems can harm already disadvantaged groups (e.g., racial minorities) (Buolamwini and Gebru 2018), for instance through spurious correlations between protected characteristics (e.g., skin colour or gender) and other features (e.g., hair length Balakrishnan et al. 2020). Therefore, methods for detecting biases in visual datasets are needed to support both critical data exploration steps when building trustworthy CV systems[1] and emerging documentation practices, such as datasheets for datasets (Gebru et al. 2021).

While algorithmic bias mitigation is well explored (Barocas et al. 2019; Ntoutsi et al. 2020; Mitchell et al. 2021; Mehrabi et al. 2021), few works tackle bias detection, especially in the CV domain. A common strategy (Fabbrizzi et al. 2022) is to extract information from the visual data, arrange it into tabular form, and quantify and mitigate bias on this data type (Zhao et al. 2017; Buolamwini and Gebru 2018; Merler et al. 2019; Wang et al. 2022). However, the reduction does not necessarily preserve all biased characteristics of the visual data space. An alternative (Kärkkäinen and Joo 2021; Steed and Caliskan 2021; Wang et al. 2022) is to extract lower-dimensional feature representations -either via hand-crafted features or pretrained deep neural networks- of visual data and measure bias for those. Unlike tabular representations, feature spaces are typically endowed with metric, topological, or vector space structures that support richer analysis[2].

In this work, we follow the last strategy and adopt the Optimal Transport (OT) theory as a mathematical framework for bias detection. In particular, in Sect. 3.2 we establish that the quadratic Wasserstein distance (W_2^2) captures deviations from a variation of demographic parity and show that its Kantorovich potentials can serve as a bias score for the individual images, where images with high potentials within a group are, on average, more distant from the other group. Hence, characteristics shared by many images with high or low potentials can be prospective sources of algorithmic bias (e.g., low-scoring images could be more likely to show hats or helmets Figure 3b), as they can be learned as proxies for the entire group. Our proposed approach is tailored to check for *selection bias* (Sections 5.1, 5.2 and 5.3), but it can also spot framing and label biases that create embedding differences (see, Sect. 2.5 for a definition of the different types of bias).

Inspired by the exploratory discrimination-aware data mining framework (Berendt and Preibusch 2014), we integrate the Kantorovich potential scoring mechanism into the bias discovery pipeline of Figure 1. There, the framework serves as a filter for the Mapper algorithm (Singh et al. 2007) to procure high-level descriptions

[1] Article 10 of the proposed AI Act of the European Commission prescribes "examination in view of possible biases" as a data governance practice for high-risk AI applications. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206. Last visited 30.01.2023.

[2] Feature extraction mechanisms can introduce additional biases, in which case -especially if they are complex- it can be difficult to differentiate between them and raw data biases. In this work we assess biases of already extracted features, irrespective of their source.

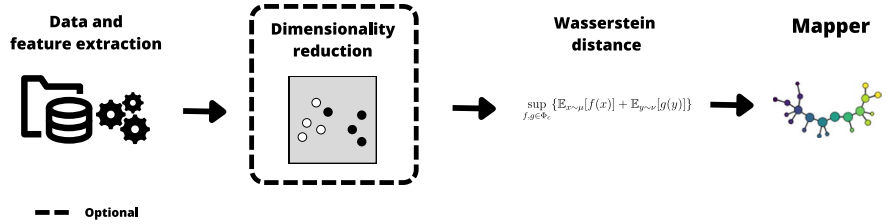


Fig. 1 The proposed bias detection pipeline (Section 4)

of CV datasets in the form of graphs. Mapper clusters images while being aware of filter function values and, by studying which are the common attributes of images in different clusters, we obtain finer-grained information on possible causes of the measured bias. The proposed pipeline guides a *qualitative* exploration of datasets and allows scientists and practitioners to formulate hypotheses on the presence of discriminatory patterns[3].

Our work presents the following contributions:

a) We set up W_2^2 as a type of binary demographic disparity in feature spaces.

b) We recognise that Kantorovich potentials can be used as a bias score for individual data samples/images.

c) We integrate these ideas into a pipeline for visual feature exploration that facilitates a qualitative high-level bias analysis of datasets.

The paper is organised as follows: Section 2 introduces required theory and reviews related approaches. Section 3 formulates the problem of quantifying biases of features extracted from a dataset, and provides theoretical guarantees for using OT as a bias detection framework. We also select and describe appropriate OT approximation (Makkuva et al. 2020) and clustering (Singh et al. 2007) algorithms that fit the framework. Section 4 demonstrates our bias investigation pipeline through a case study on gender biases in the CelebA dataset (Liu et al. 2015). Section 5 presents an experimental evaluation of theoretical properties. Finally, Sects. 6 and 7 provide a discussion on the merit and caveats of our work, as well as an outline of future research directions. The reader can find our code at https://github.com/sfabbrizzi/OT-ICNN-bias.

[3] Discrimination has a precise legal meaning, namely disparities on the ground of non-acceptable features. Since our work discovers many types of disparities, we describe it as *bias-aware data mining*. Practitioners should decide which disparities pertain to actual discrimination.

2 Background

In this section we provide theoretical concepts needed to understand our analysis, and overview bias exploration in CV and related attempts to apply OT in bias detection.

2.1 Notation

Throughout our analysis, we work with a feature extractor $F : \mathcal{I} \to \mathbb{R}^d$, where $\mathcal{I}$ is an image dataset of which the samples are converted to d-dimensional feature spaces. In these datasets, we consider a sensitive binary attribute S that assumes values among $\{s_0, s_1\}$ for each data sample. We also denote the sample distributions of images sharing sensitive attribute value $S = s_0$ and $S = s_1$ as μ and ν respectively (e.g., women and men). That is:

$$\mu = \frac{1}{|\mathcal{I}_{s_0}|} \sum_{x_i \in \mathcal{I}_{s_0}} \delta_{x_i} \text{ and } \nu = \frac{1}{|\mathcal{I}_{s_1}|} \sum_{x_i \in \mathcal{I}_{s_1}} \delta_{x_i}$$

where $\mathcal{I}_s = \{x \in \mathcal{I} \mid x \text{ has sensitive attribute value} S = s\}$ and δ_x is the Dirac distribution centered at x, which lets us move from integral definitions to discrete computations.

2.2 Optimal transport

Given two probability measures μ and ν over two spaces X and Y and a (lower semi-continuous) cost function $c : X \times Y \to \mathbb{R}$, OT theory (Villani 2003, 2008) searches for the minimal cost of transferring one measure into the other. In this work, the probability measures correspond to the distribution of samples with different sensitive attribute values, for example men and women, in the same image feature space (in that case $X = Y$).

As an intuitive formulation for the OT theory (Villani 2003), one can imagine a set of mines and factories, and the respective distributions of coal production and demand. In this setting, solving an OT problem would mean finding the most cost-effective transport map of which mine should provide coal to which factory. This translates to the following objective, known as Monge's problem:

$$\inf_{T:X \to Y} \int_X c(x, T(x)) d\mu \tag{1}$$

where $\nu = T_{\#}\mu$ and $T_{\#}$ is the push-forward of μ along the function $T : X \to Y$.

The condition of T being a function can be too restrictive in that it transfers the entire production of each mine to only one factory, thus creating a potentially insolvable Monge's problem (Villani 2008). To address this theoretical concern, the more general Kantorovich's problem aims to find the coupling between two

measures μ and ν that allows the masses allocated on individual elements to be split. This is formulated per the following objective:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y) d\pi \tag{2}$$

where $\Pi(\mu, \nu)$ denotes set of couplings between μ and ν.

When the sets are the same d-dimensional Euclidean spaces $X = Y = \mathbb{R}^d$, one popular cost function is the quadratic Euclidean distance $c(x, y) = \frac{1}{2}\|x - y\|_2^2$. Under these circumstances, the square root of the solution to Equation 2 is a metric in the space of measures (with finite second moments) over $\mathbb{R}^d$ and is usually referred to as the 2-Wasserstein distance $W_2(\mu, \nu)$. This property is the one that guarantees that Wasserstein distance is suitable for checking demographic parity (Sects. 2.3 and 3.2).

Solving the OT problem in the space of couplings is usually difficult, but it is equivalent to the tractable dual form (Villani 2003):

$$W_2(\mu,\nu)^2 = \sup_{(f,g)\in\Phi_c} \int_X f(x) d\mu + \int_Y g(y) d\nu \tag{3}$$

where W_2^2 is the quadratic Wasserstein distance, Φ_c is the space of function pairs $(f, g) \in L^1(d\mu) \times L^1(d\nu)$ such that $f(x) + g(y) \leq \frac{1}{2}\|x - y\|_2^2$ for $d\mu$-almost all $x \in X$ and $d\nu$-almost all $y \in Y$. An intuitive interpretation (Villani 2003) of the dual form is that, instead of minimising the transportation cost of coal from the mines to factories, it maximises the profit of an external company to which the transportation problem is outsourced. The values of $f(x)$ and $g(y)$ respectively define the company's price for loading the coal at a mine x and unloading it at the factory y and are known as *Kantorovich potentials*.

We state two results (Villani 2003) that are useful in the remainder of this work. Given that the probability measures μ and ν over $\mathbb{R}^d$ that exhibit finite second-order moments the following hold true:

Knott/Smith optimality criterion. A coupling $\pi \in \Pi(\mu, \nu)$ is optimal for Equation 2 under the quadratic cost function if and only if there exists a convex lower semi-continuous function ϕ such that for $d\pi$-almost all (x, y) it holds that $y \in \partial\phi(x)$, with $\partial f(x)$ being the subgradient of f in x. Furthermore, the couple (ϕ, ϕ^*) is a solution of Equation 3, where ϕ^* is the convex conjugate of ϕ.

Brenier's theorem. If ν is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, there exists a unique optimal coupling $\pi = (\nabla\phi^* \times Id)_{\#}\nu$. Furthermore, $\nabla\phi^*$ is the unique solution of Monge's problem.

2.3 Demographic parity

Fairness definitions can be categorised (Mehrabi et al. 2021) as either group, subgroup, or individual fairness. The first two aim to balance properties of interest, such

as the percentage of positive predictions or misclassification rates, between groups and random subgroups respectively. Individual fairness (Dwork et al. 2012) aims to treat similar individuals similarly.

In this work, we capture distributional differences between two groups of images, which is a type of group fairness. In particular, this consideration is a form of demographic parity (Kamiran et al. 2010; Kamiran and Calders 2012; Kamishima et al. 2012; Ntoutsi et al. 2020), which stipulates that both groups should exhibit similar statistical properties. Demographic parity typically involves checking for proportional representation of each group in outputs of AI systems. In this work's terminology, demographic parity can be written as:

$$\mu(F^{-1}(1)) = \nu(F^{-1}(1)) \tag{4}$$

where, for this definition, $F : X \to \{0, 1\}$ is a binary classifier. In this case, this is equivalent to the otherwise stronger condition[4]

$$F_{\#}\mu = F_{\#}\nu \tag{5}$$

Contrary to other types of group fairness, such as disparate mistreatment that aims to create similar error rates between groups (Zafar et al. 2017), demographic parity can also be quantified on reference or training data by measuring the representation of groups in desired prediction outcomes instead. As such, it has been used to quantify raw dataset biases as well (Sattigeri et al. 2019; Ding et al. 2021).

In this work we have a similar goal of quantifying CV dataset bias, although we are looking to do so regardless of downstream tasks in which datasets are used. Thus, we adopt Equation 5 as a fairness definition with the only change that $F : X \to \mathbb{R}^d$ is an embedding into a high dimensional Euclidean space. The metric properties of the Wasserstein distance ensure that it is a good way to measure demographic parity in feature spaces.

2.4 OT and bias

OT theory has recently been used as fairness constraints to enforce variations of demographic parity in bias mitigation tasks (Jiang et al. 2019; Gordaliza et al. 2019; Zehlike et al. 2020; Chiappa et al. 2020; Chiappa and Pacchiano 2021), to measure and explain bias in classification or (1-dimensional) regression (Miroshnikov et al. 2022) or generally look at the problem of bias through the lens of OT (Kwegyir-Aggrey et al. 2021).

Most OT bias works intend to understand and mitigate model bias and are not suited to our data mining and bias discovery perspective. The approach closest to ours is the one of Kwegyir-Aggrey et al. (2021); it introduces both an individual and a global measure of bias, where *individual bias* refers to individual sample bias and is defined as:

[4] For binary classifiers $F_{\#}\mu(1) = \mu(F^{-1}(1)) = 1 - F_{\#}\mu(0)$. For continuous or multi-dimensional values, this condition requires equal distribution between groups.

$$u(x_0) = \int_Y c(x_0, y) d\pi(x_o, \cdot) = \int_Y c(x_0, y) d\nu \tag{6}$$

where the cost function c is the distance $d(x_0, y)$. Later, in Proposition 1, we show that the above rankings and Kantorovich potentials are equivalent when the cost function c is the *square* distance. This approach also sums all individual bias scores to procure a global measure of bias, which is a coarser measure than W_2^2 in that it does not take the transport map into account. Finally, this approach was originally devised to measure bias in the prediction space of tabular data, instead of the feature space of vision data.

2.5 Bias in visual data

Fabbrizzi et al. (2022) found that bias in visual data can be roughly divided into three categories: *selection bias*, *framing bias*, and *label bias*. Selection bias refers to "disparities or associations created as a result of the process by which subjects are included in a visual dataset". Framing bias comprises "associations or disparities that convey different messages and/or can be traced back to the way in which the visual content has been composed". Finally, label bias comprises "errors in the labelling of visual data, either with respect to some ground truth, or due to use of poorly defined or inappropriate semantic categories".

Previous findings indicate that the presence of visual data biases can indeed be detected early. Characteristically, Torralba and Efros (2011) showed that benchmark datasets for object detection tend to exhibit heavy selection biases. For example, out of popular vision datasets that include depictions of cars, ImageNet (Deng et al. 2009) exhibits a strong preference for racing cars, while Caltech101 (Li et al. 2022) for side views of cars. These biases were found by training an SVM (Cortes and Vapnik 1995) to distinguish the source of images. The detection accuracy of the model serves as a global indication of how biased the datasets were. The same method can check for bias *within* a dataset, by training the SVM to distinguish between groups of samples.

The confidence of SVM predictions can help us draw conclusions for individual samples, where high confidences indicate more biased images, i.e., which exhibit characteristics intrinsic to datasets or groups of samples. While this method was introduced as a toy experiment, it remains popular in the literature (Tommasi et al. 2015; Panda et al. 2018; Kärkkäinen and Joo 2021; Wang et al. 2022). In Sect. 5.1, we employ a variation of SVM-based biased image detection as a baseline against which we compare our approach.

3 Optimal transport for bias detection

In this section, we detail the process through which OT theory becomes a useful tool for bias detection. First, Sect. 3.1, describes the W_2^2 approximation method of (Makkuva et al. 2020) and why it is suited to our needs. Second, Sects. 3.2 and 3.3

show why the approximated distance and the Kantorovich potentials form an appropriate framework for bias-aware data analysis. Finally, Sect. 3.4 describes a clustering algorithm (Mapper by Singh et al. (2007)) that allows a finer-grained qualitative analysis of data based on potentials.

3.1 Computationally tractable OT approximation

The dual formulation of the Kantorovich's problem in Equation 3 lets us compute W_2^2 by approximating the Kantorovich potentials f and g with deep neural networks (Makkuva et al. 2020). We also follow this technique, because it works around the expensive computational demands of solving OT problems for high-dimensional data (Cuturi 2013) by considering the following reformulation of the Kantorovich dual problem:

Theorem 1 (Makkuva et al. 2020) *Given two probability distributions μ and ν with the latter being absolutely continuous with respect to the Lebesgue measure, then*

$$W_2(\mu,\nu)^2 = C_{\mu,\nu} + \sup_{\substack{\phi \in \texttt{CVX}(\mu) \\ \phi^* \in L^1(\nu)}} \inf_{\psi \in \texttt{CVX}(\nu)} \mathcal{V}_{\mu,\nu}(\phi,\psi) \tag{7}$$

where $C_{\mu,\nu} = \frac{1}{2}\left(\int_X \|x\|_2^2 d\mu + \int_Y \|y\|_2^2 d\nu\right)$, $\mathcal{V}_{\mu,\nu}(\phi,\psi) = -\int_X \phi(x)d\mu - \int_Y \langle y, \nabla\psi(y)\rangle - \phi(\nabla\psi(y))d\nu$, *and* CVX *represents the set of convex functions.*

This reformulation enables approximation of W_2^2 by substituting $\texttt{CVX}(\mu)$ and $\texttt{CVX}(\nu)$ with the set $\texttt{ICNN}(\mathbb{R})$ of scalar-valued Input Convex Neural Networks introduced by Amos et al. (2016). A byproduct of solving this problem is the pair of convex functions ϕ and ψ, which can be used to express the Kantorovich potentials as $f(x) = \frac{1}{2}\|x\|_2^2 - \phi(x)$ and $g(y) = \frac{1}{2}\|y\|_2^2 - \psi(y)$ for f and g as in Kantorovich duality of Equation 3. Hence, we easily compute the value of f for image features and rank images according to the produced potentials.

3.2 W_2^2 to quantify dataset bias

The push-forward formulation of demographic parity at the end of Sect. 2.3 is also applicable to high-dimensional systems, such as image feature vectors outputted by feature extractors. For any (Borel) subset of such feature spaces, the formulation imposes that the probability of sampling the feature vector of an image from such a subset does not depend on the demographic attribute. Note that this property is stronger than comparing distributions through their average. A visual demonstration can be found in Figure 2.

As the Wasserstein distance is a metric on the space of probability measures with finite second moments, we have that $F_\#\mu = F_\#\nu$ iff $W_2(F_\#\mu, F_\#\nu) = 0$. Thus, we propose using W_2^2 as a measure of bias:

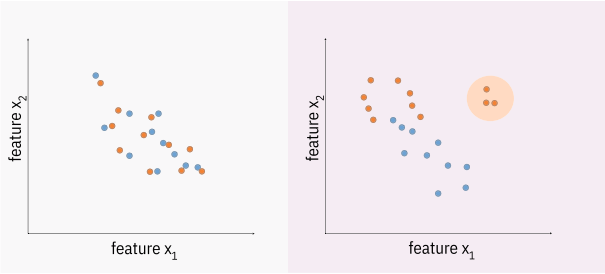


Fig. 2 On the left, features are distributed in their space without being affected by the attribute of interest (e.g, gender depicted in blue/orange). On the right, a feature extractor exhibits biases with respect to that attribute and some orange samples form a special sub-group (highlighted) that merits further investigation

Definition 1 Consider an image dataset $\mathcal{I}$, a feature extractor $F : \mathcal{I} \to \mathbb{R}^d$, and μ and ν the distributions of the images depicting two values of a protected attribute. We say that $\mathcal{I}$ is F-biased if $W_2(F_{\#}\mu, F_{\#}\nu) \neq 0$ where $F_{\#}\mu$ and $F_{\#}\nu$ are the push-forward distributions along the feature extractor F of μ and ν on $\mathcal{I}$.

3.3 Kantorovich potentials to score image bias

While we used W_2^2 to assess bias at the dataset level, we also explore the contribution of images with features $x_i \in \mathcal{I}$ drawn from μ by capturing how far they are from the whole distribution ν. To do this, we look at the interpretation of the dual Kantorovich problem in Equation 3; Kantorovich potentials $f(x_i)$ compute the prices for moving x_i into their correspondent image sampled from ν to maximise profit. Intuitively, the greater the distance the higher the price, at least on average. We theoretically express this property below:

Proposition 1 *Let $\mu = \sum_{i=0}^{n} a_i \delta_{x_i}$ be a strictly positive discrete probability measure and ν be an absolutely continuous probability measure. Given the Kantorovich potentials f and g, solutions to the problem 3, we have that for every $x_0, x_1 \in X$ if $f(x_0) \geq f(x_1)$ then $\int_Y c(x_0, y)d\nu \geq \int_Y c(x_1, y)d\nu$ and vice versa.*

Proof $f(x_0) \geq f(x_1)$ implies that for every $y \in Y$ $f(x_0) + g(y) \geq f(x_1) + g(y)$. Being f and g solutions to the Kantorovich dual problem in Equation 3, by Knott-Smith optimality criterion in Sect. 2.2, we have for all $x \in X$ and for almost all $y \in Y$ that $f(x) + g(y) = c(x, y)$. Note that the property holds for all $x \in X$ and not for *almost all* $x \in X$ because the only measure-zero set for a strictly positive discrete measure is the empty set. The result follows by integrating over Y.

Vice versa, $\int_Y c(x_0, y)d\nu \geq \int_Y c(x_1, y)d\nu$ implies that $\int_Y f(x_0) + g(y)d\nu \geq \int_Y f(x_1) + g(y)d\nu$. Thus, $f(x_0) + \int_Y g(y)d\nu \geq f(x_1) + \int_Y g(y)d\nu$ which, in turn, implies $f(x_0) \geq f(x_1)$. □

Given the above result, we propose using the potential function f to rank the images in X according to how costly they are to be transported into the distribution ν. This can drive qualitative bias analysis that inspects higher-ranking images for common characteristics. This is fundamental to understanding whether the differences in the distributions of two protected attributes' values are due to discriminatory characteristics or not.

The approximation method reviewed in Sect. 3.1 works under the hypothesis of Brenier's theorem, which stipulates that at least one of the two probability distributions is absolutely continuous. Since we end up with two sample feature distributions, we adjusted the algorithm's original form presented by Makkuva et al. (2020) to ensure that the theorem's requirement is met. The adjustment consists of sampling from the kernel density estimation of one of the two distributions instead of directly parsing raw data.

3.4 The Mapper algorithm

The Mapper algorithm (Singh et al. 2007) is a method for reducing high-dimensional data to a graph, called the *nerve complex*, that provides a high-level understanding of the topological structure of the data. We want to exploit Mapper's properties to get a fine-grained bias analysis of the data.

In particular, assume a dataset D lying in a topological space X and a continuous function $f : X \to \mathbb{R}$ that quantifies a certain property of the data to be studied. Given a covering $\mathcal{U}$ of $\mathbb{R}$, Mapper stipulates that a covering of X is obtained from the pre-image $\mathcal{V} = \{V \subseteq f^{-1}(\mathcal{U}) \text{ connected components}\}$. Thus, it defines the nerve complex graph $N(\mathcal{V})$ of the covering by considering as nodes the connected components of the sets $\mathcal{V} = f^{-1}(\mathcal{U})$ and forming edges between V_i and V_j only if $V_i \cap V_j \neq \emptyset$.

To mimic this graph construction process on a dataset D, whose discrete data sample nature prevents the identification of connected components, Mapper performs a clustering step on the pre-image space of each $U_i \in \mathcal{U}$ along $f|_D : D \subseteq X \to \mathbb{R}$. Thus, the clusters play the role of connected components.

Mapper works with any clustering algorithm. This way, we obtain a clustering of D and the relative nerve complex $N(\mathcal{V}|_D)$. Depending on the filter function f, the clustering captures different information that we can use for exploratory data analysis. This makes Mapper a suitable choice for our framework as it allows clustering the data using the Kantorovich potentials introduced in Sect. 2.2 as a filter function.

4 A case study on CelebA gender bias

This section presents an application of our framework on CV data bias exploration. This case study fleshes out the pipeline of Figure 1 with predetermined feature extractors, but can be also used in combination with other extractors or datasets. The pipeline's usefulness is verified both empirically and with the aid of a user study. Refer to Appendix B for implementation details.

4.1 Data and feature extraction

Our study is conducted on the well-known benchmark CelebA dataset (Liu et al. 2015). This contains 202,600 face images, each of which is supplied with 40 different attributes. The dataset is derived from the benchmark CelebFaces+ (Sun et al. 2014) by annotating the latter with the help of a professional data annotation company. The annotation includes information about the gender of subjects depicted in each image.

We follow the pipeline outlined in Figure 1, that aims to detect bias *after* feature extraction, but before training any system. In principle, our OT results could be applied to any kind of feature spaces, as long as these are endowed with a metric. However, the adopted approximation method of Makkuva et al. (2020) (Sect. 3.1) is tailored to Euclidean spaces only and, thereon, deep-neural network features.

Our approach admits any (Euclidean) feature extractor and we explore three popular ones: a) *ResNet* (He et al. 2016) is a well-known convolutional neural network architecture used for a variety of vision tasks (e.g., object classification, object recognition, segmentation). We use its PyTorch implementation[5] with the default weights pre-trained on ImageNet (Deng et al. 2009) as a feature extractor by cutting out its fully connected output layer. b) *Autoencoder* features trained with a CNN-based encoder and decoder to minimise the reconstruction loss of CelebA dataset. c) *FaceNet* (Schroff et al. 2015) is an embedding framework used for facial recognition and face images clustering. We used a Pytorch implementation[6] with default weights pre-trained on VGG Faces 2 (Cao et al. 2018).

4.2 Dimensionality reduction (optional)

Clustering methods in the last step of our pipeline are often affected negatively by high-dimensionality. Here, we support that step by preemptively performing dimensionality reduction. This step is not always necessary, and may (should) be omitted for exploration of low-dimensional features, where it risks impacting the image ranking induced by Kantorovich potentials. Hence, it is important to select a dimensionality reduction approach (e.g., among popular ones) that mostly preserves ranking order. We show what this selection entails in Sect. 5.4, which leads us to reduce the extracted 512-dimensional features to 3 dimensions with PCA. This reduction mostly preserves image ranks produced by the next step while also creating clustering-friendly image representations.

4.3 Compute Kantorovich potentials

To score -and ultimately rank- images based on their contribution to demographic disparity, we approximate W_2^2 in the way described in Sect. 3.1, which produces

[5] https://pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html. last visited 05.12.2022.

[6] https://github.com/timesler/facenet-pytorch. Last visited 05.12.2022.

(a) Top 18 for ResNet18. (b) Bottom 18 for ResNet18.

(c) Top 18 for autoencoder. (d) Bottom 18 for autoencoder.

(e) Top 18 for FaceNet. (f) Bottom 18 for FaceNet.

Fig. 3 Images of women in CelebA with the top (left) and bottom (right) 18 Kantorovich potentials for ResNet18, autoencoder, and FaceNet features

Kantorovich potentials as a computational byproduct. We obtain the following approximations to serve as bias scores for different feature extractors: 20.97 for ResNet, 29.88 for the autoencoder, and 0.06 for FaceNet. If embedding spaces are the same (e.g., are all $\mathbb{R}^{512}$), such scores become comparable. If some feature extractor shrinks the space (for example, FaceNet features are normalised), we are going to obtain smaller bias scores, without this necessarily implying lesser bias.

As stipulated in Sect. 3.2, we retain the Kantorovich potentials during computations and use them to rank women in the dataset. Then, in Figure 3, we show the images of women with the top and bottom 18 potentials for each of the feature extraction methods. The most common attribute among the top images for ResNet18

and autoencoder is having long hair, while women in the bottom images either have their hair tied up or shorter hair, or wear some sort of headgear. This indicates that hair length could be a discriminative factor when correlated with predictive tasks and should therefore be kept in mind when using the first two types of features to produce new systems. To make matters worse, the ranking for the autoencoder feature seems to also be correlated with skin colour and this warrants additional investigation. On the contrary, a visual inspection of the most extreme ranking for FaceNet features does not highlight apparent prospective sources of discrimination for these.

4.4 Mapper visualisation

We finally conduct a finer-grained analysis throughout the available data. To do this, we follow the methodology suggested in Sect. 3.4 and apply the Mapper algorithm using the Kantorovich potential functions as filters; the resulting nerve complex guides our bias investigation. To run the algorithm, we enlist K-Means as its clustering component and select a number of clusters for each of its layers by applying the elbow technique on the clustering inertia measure (Cui et al. 2020). This retains a homogeneous level of Kantorovich potentials within each cluster.

Figure 4 shows the nerve graph for ResNet18 features produced by Mapper. We uniformly sample images from the clusters that compose the shortest path between two clusters in the first and last layers. As hypothesised from the ranking inspection, women in the lower-layer clusters (purple/blue nodes in the nerve graph) have their hair covered and at intermediate layers have progressively longer and less covered hair, until the end-result is more pronounced in the top layers (yellow nodes). This particular path visually corroborates the previous step's hypothesis that hair length and obfuscation is a characteristic captured by ResNet18, and hence should be kept in mind as a potential source of bias.

Similar behaviour is exhibited by the autoencoder's features, where once again hair length increases as we move from the lower layer to the higher ones. The full exploration can be found in Appendix C. Previous work (Balakrishnan et al. 2020) points out that the hair length characteristic might be correlated with skin colour and would turn an apparently innocuous bias into something potentially discriminatory. As for the FaceNet features, the Mapper analysis does not provide any additional interesting information about possible biases.

4.5 User study

To address potential confirmation biases during the above investigation, we also set up a complementary user study. This aims to check whether other people can also discover prospective biases in CelebA images by following our approach compared to randomised dataset exploration. The study gathered responses by 25 participants from the authors' research groups, which viewed image batches and were asked to answer questions pertaining to bias insights they revealed about the data. 13 of the participants were randomly assigned as a control group and were shown uniformly sampled images, whereas the rest were shown batches produced

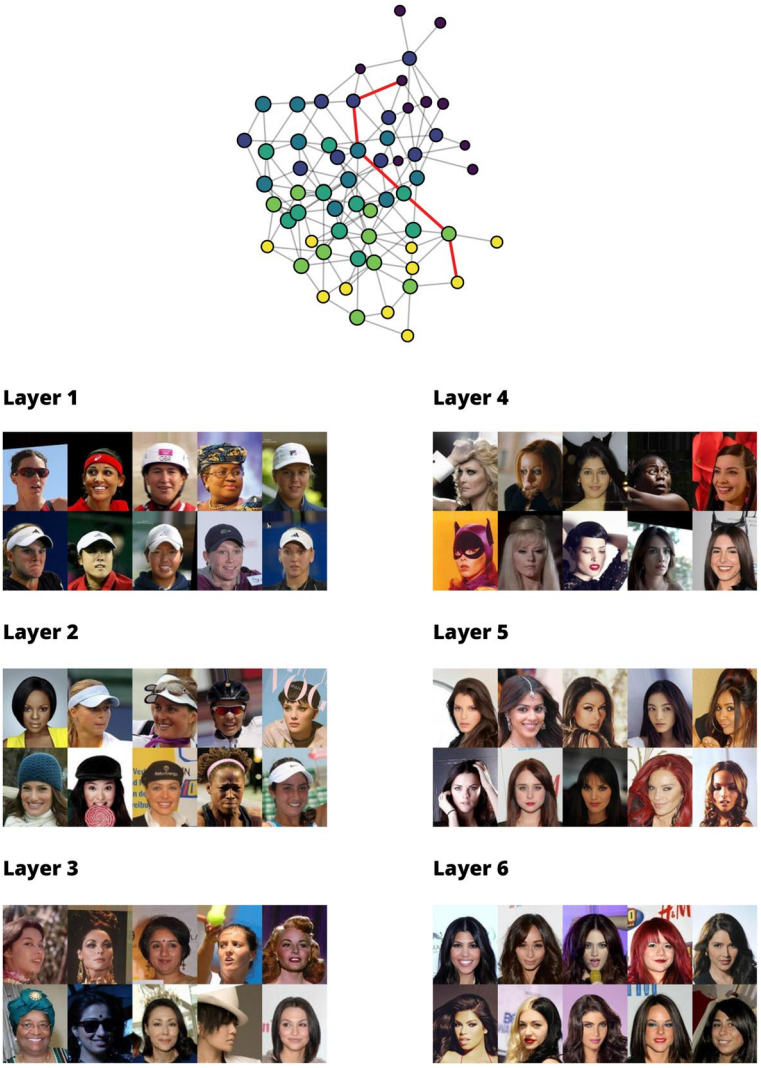


Fig. 4 A path in the nerve complex constructed by Mapper for the three-dimensional ResNet18. In purple we have the bottom scoring images (layer 1), in yellow the top scoring ones (layer 6)

with our methodology. Further details on the questions and their analysis can be found in Appendix D. Overall, we found that participants in the experiment group would be able to find a variety of biases in the batches we showed them and, in particular, to find a different representation of women between the images ranking high with respect to the Kantorovich potentials and those ranking low. On the contrary, the participants in the control group were prone to find biases that are very prominent in CelebA (e.g., majority of white women) and when asked to compare two batches they would be less confident in describing one of them

as more biased. Generally, we draw the conclusion that, while there is some possibilities to fall into some confirmation bias, our method is useful to formulate hypotheses about possible biases in the data.

5 Experiments

The previous section showcases how our framework can discover biased visual features in the CelebA dataset. In this section, we provide experimental evidence to back theoretical claims supporting our pipeline. To begin with, in Sect. 5.1, we compare W_2^2 with two baselines. Then, in Sects. 5.2 and 5.3, we assert that we can capture bias both for whole datasets and for individual images. Finally, in Sect. 5.4, we assess the impact of dimensionality reduction on W_2^2 and the image ranking based on Kantorovich potentials; we show that several reduction techniques retain minimal impact and therefore can be used as part of Sect. 4's pipeline to facilitate the usage of Mapper. A discussion on our findings is conducted in Sect. 6. Refer to Appendix B for implementation details.

5.1 Baseline comparison

Goal. The notion of statistical parity led us to select W_2^2 as a measure of disparity. We investigate whether this choice can quantify dataset bias as well as alternatives, like the popular SVM-based approach of Torralba and Efros (2011) and the measures proposed by Kwegyir-Aggrey et al. (2021).

Torralba and Efros (2011) measure how separable two datasets or groups of samples are. Analogously to our approach, they set SVM accuracy as a global measure of bias, and classification confidence for each image as a bias contribution score. To make a fair comparison between this approach and our use of W_2^2 , we replace accuracy (which suffers from the limitation of being bounded in [0, 1]) with the average quadratic distance of points from the SVM's decision boundary $\mathcal{B}$:

$$\frac{1}{|X|+|Y|} \sum_{p \in X \sqcup Y} \frac{1}{2} d(p, \mathcal{B})^2 \tag{8}$$

This way, the distance $d(p, \mathcal{B})$ replaces the (lack of) classifier confidence.

We also consider a variation of this methodology that replaces the SVM with a fully connected neural network. Computing the distance from the latter's decision boundary is not as straightforward, and we approximate it by repurposing the LIME algorithm (Ribeiro et al. 2016):

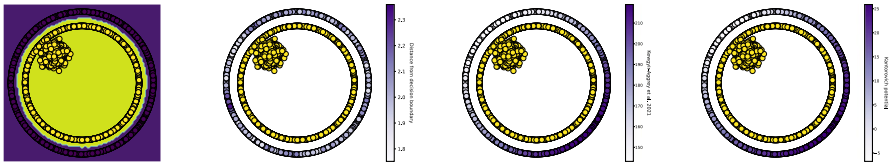


Fig. 5 (**a**) Decision boundary of a 10-layer neural network with perfect accuracy. (**b**) The distance from the boundary of the outer circle's points, which does not depend in the inner distribution. (**c**) Individual bias score as presented in Kwegyir-Aggrey et al. (2021) (**d**) Kantorovich potentials for the outer circle, which are lower near the inner Gaussian

Algorithm 1 Approximate distance from the decision boundary.

Input: Model f, point $x \in X$, Kernel function $\mathcal{K}$.
Output: Approximated $d(x, \mathcal{B}_f)$ or $\frac{1}{2} \cdot d(x, \mathcal{B}_f)^2$.

1: Sample $Y \sim \mathcal{N}(x, \Sigma)$
2: Compute the predictions $f(Y)$
3: Weight the points in Y according to the kernel function $\mathcal{K}$
4: Train a weighted linear SVM g on $(Y, f(Y))$
5: Return $d(x, \mathcal{B}_g)$ or $\frac{1}{2} \cdot d(x, \mathcal{B}_g)^2$

Kwegyir-Aggrey et al. (2021) instead propose a linear individual bias measure that depends on averaging over pairwise sample distances, as shown in Equation 6. To make this comparable to our approach, we create an adjusted variation that uses the same squared distance formulation as W_2^2 , i.e., $c(x_0, y) = \frac{1}{2} d(x_0, y)^2$. As a global bias measure, this approach uses the sum of individual biases, which depends on the number of samples. We again make this comparable to our approach by averaging individual biases.

Setup. To show that there are cases in which W_2^2 is more expressive than model-based baselines in quantifying bias, we crafted two non-linearly-separable synthetic datasets, for which SVMs tend to exhibit diminishing learning power. According to Proposition 1 the approach of Kwegyir-Aggrey et al. (2021) would create the same ranking as the Kantorovich potentials, but the global bias measure would be coarser due to averaging all transportation costs instead of taking into account the OT map.

We create the datasets by uniformly sampling points from two concentric circles in a two-dimensional feature space, where each circle corresponds to a different sensitive attribute value. In the first dataset, the circles are perfectly concentric with radii 10 and 7.5. In the second case, we added points to the inner circle sampled from a Gaussian distribution centred in $(-4, 4)$ and covariance matrix $0.5 \cdot I_2$. In these settings, the distance from SVM decision boundaries does not provide any information about the two distributions; but even a more versatile learning model, like a neural network, could leave some biases undetected.

Results. For the first synthetic dataset, we compare a) W_2^2 , b) the measure proposed by Kwegyir-Aggrey et al. (2021), and the quadratic distance from the decision boundary of both a c) linear SVM and d) our variation with a 10-layer neural

network classifier. For these approaches, we respectively obtained dataset bias scores of 1.98, 82.0, 21.05, and 1.22.

For the second synthetic dataset, we compute the distance from the decision boundary of a deep neural network with the Kantorovich potentials on the dataset's outer circle. Figure 5 demonstrates that, even for a perfectly accurate decision boundary, the distance is evenly distributed across the outer circle and can therefore be a misleading indication of individual bias. On the other hand, Kantorovich potentials exhibit higher values on the side opposite to the Gaussian blob, and therefore correctly identify that points on that side differ more from the inner distribution than those near the blob. We also verify that the result for Kwegyir-Aggrey et al. (2021) is nearly-identical to the Kantorovich potentials, up to the goodness of our approximation.

5.2 W_2^2 captures total dataset bias

Goal. We now investigate the ability of W_2^2 to capture different degrees of data bias under the same feature extractors.

Setup. We experiment with two datasets: CelebA and Biased MNIST (Shrestha et al. 2022). We construct splits of CelebA with different degrees of selection bias, that is, we sample four sub-datasets ensuring that respectively 90%, 60%, 30%, and 10% of the samples have a positive value for a specified binary attribute (e.g, wearing ties, hats, or eyeglasses). W_2^2 is computed on the 12 datasets that combine different attributes and selection biases, as well as for a -this time- uniformly sampled dataset of the same size of 9K images (in the latter, 7% of people wore ties, 7% wore glasses, and 4% wore hats). In Biased MNIST, we compare its training splits (50K images) for available bias levels (0.1, 0.5, 0.75, 0.9, 0.95, and 0.99) with the dataset's test split (10K images), which is unbiased. For both datasets, we experiment with the ResNet18 feature extractor.

Results. Figure 6 shows that W_2^2 monotonically reflects the degree of artificially injected bias in most cases, except for the smiling attribute.

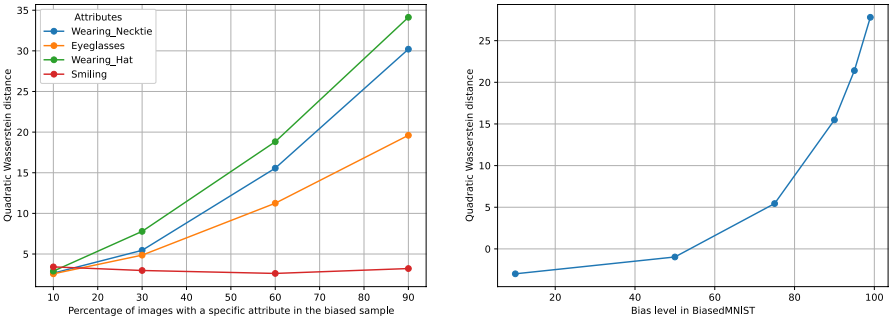
5.3 Kantorovich potentials capture image bias

5.3.1 Goal

We next investigate whether Kantorovich potentials capture the contribution of images to the total dataset bias. To this end, we assess whether the distribution of potentials computed on biased samples is skewed towards higher values for images with attribute values capturing biased characteristics.

5.3.2 Setup

We follow the same setup as in the previous experiment.



(a) CelebA. (b) Biased MNIST.

Fig. 6 Changes in the approximate W_2^2 over the increment of selection bias across (**a**) four CelebA attributes and (**b**) the Biased MNIST dataset

5.3.3 Results

A two-sample Kolmogorov-Smirnov (KS) test checks whether images with the artificially biased attribute (e.g, Wearing_Necktie) exhibit higher Kantorovich potentials than others for every cumulative distribution threshold of the potentials. Tables 1 and 2 show test outcomes across several rates of selection bias.

The high p values in Table 1 indicate that the cumulative function distribution of Kantorovich potentials for the images with the attributes in the first column are higher than the same distributions for images without the selected attributes in

Table 1 KS p values rounded to three decimal places for CelebA

Attribute	10%	30%	60%	90%
Wearing_Necktie	0.953	1	1	1
Eyeglasses	1	1	1	1
Wearing_Hat	1	1	1	1
Smiling	0	0	0.001	0.997

Table 2 KS p values rounded to three decimal places for the digit 4 in Biased MNIST

Selection bias	10%	50%	75%	90%	95%	99%
digit_color_ix	0.268	0.318	0.234	0.018	0.568	0.568
digit_position_ix	0.052	0.828	0.264	0.860	0.689	0.649
digit_scale_ix	0.052	0.828	0.264	0.860	0.689	0.649
letter_color_ix	0.503	0.806	0.632	0.332	0.042	0.471
letter_ix	0.619	0.653	0.342	0.118	0.064	0.621
texture_color_ix	0.337	0.902	0.222	0.162	0.297	0.456
texture_ix	0.291	0.942	0.422	0.465	0.735	0.882

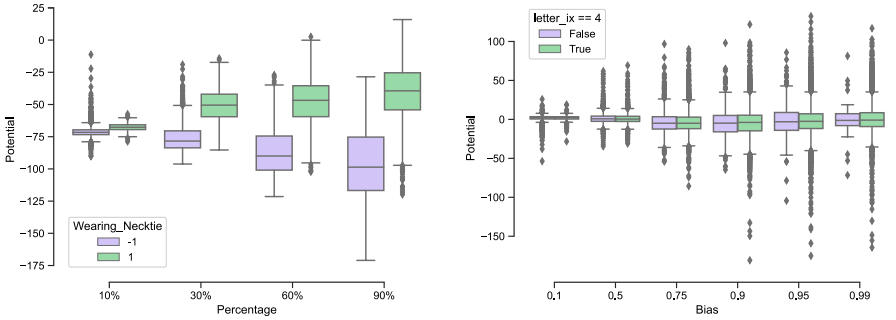


Fig. 7 Kantorovich potential box plots for the CelebA attribute Wearing_Necktie (left) and for the Biased MNIST digit 4 (right) with different degrees of selection bias

CelebA. On the other hand, worse confidences are exhibited in Table 2, which leave us uncertain with regards to our hypothesis.

To get a sense of what confident bias assessment entails, in Figure 7 we show differences in Kantorovich potential distributions as more selection bias is introduced for the Wearing_Necktie attribute in CelebA images. Similar figures can be found for other explored attributes of this dataset in Appendix E. As already indicated by the KS test, differences between the distribution of the Kantorovich potentials with respect to the attribute of interest grow more pronounced for more biased data (this is the desired behaviour). Furthermore, distributions are mostly skewed towards higher values for the images with the attribute. Overall, we confirm a general trend for Kantorovich potentials to capture the artificially injected data bias.

For Biased MNIST, W_2^2 captures the overall increment of bias pretty well (Figure 6b); however, if we compute the distribution of Kantorovich potentials for each digit with respect to the attributes that are correlated with that digit, we do not observe the same trend. See, for example, Figure 7. This is a surprising behaviour that certainly needs further exploration.

5.4 Impact of dimensionality reduction

5.4.1 Goal

When dealing with image data in the context of our framework, the high dimension of the extracted features can negatively impact the clustering required to run the Mapper algorithm. Here, we investigate the impact of dimensionality reduction techniques on our approach in terms of: a) to what extent such techniques affect the approximation of the quadratic Wasserstein distance, and b) whether the ranking defined by Kantorovich potentials changes after dimensionality reduction.

5.4.2 Setup

We experiment with four popular dimensionality reduction techniques; PCA (Hotelling 1933), Isomap (Tenenbaum et al. 2000), Spectral Embedding (SE by Belkin and

Table 3 W_2^2, and the Spearman correlation r between Kantorovich potentials computed before and after reduction for different percentages of people wearing hats in the biased sample. All p values are less than 10^{-39}

Wearing hats	10%		30%		60%		90%	
	W_2^2	r	W_2^2	r	W_2^2	r	W_2^2	r
PCA 3	0.067	0.535	2.244	0.959	8.615	0.966	18.378	0.930
PCA 50	0.651	0.559	4.039	**0.973**	13.509	0.989	27.048	0.984
PCA 150	2.196	0.615	6.000	0.967	16.550	0.994	31.129	0.993
Isomap 3	2.256	0.371	47.547	0.847	169.607	0.887	349.713	0.807
Isomap 50	6.269	0.355	86.104	0.890	299.818	0.921	603.285	0.852
Isomap 150	32.744	0.408	138.202	0.909	403.700	0.939	779.033	0.878
SE 3	1.192	-0.139	1.197	-0.404	1.187	0.679	1.211	0.243
SE 50	0.029	-0.404	0.029	-0.402	0.029	-0.390	0.029	-0.376
SE 150	0.080	-0.478	0.080	-0.475	0.080	-0.389	0.080	-0.310
t-SNE 3	1.791	0.331	32.032	0.853	131.591	0.869	250.732	0.733
Not reduced	2.894	1	7.785	1	18.818	1	34.127	1

Best correlation in bold

Niyogi (2003)), and t-SNE (Van der Maaten and Hinton 2008). We extract ResNet18 features of four CelebA splits composed by joining the uniform sample and the sample with 10%, 30%, 60% and 90% of people wearing hats as described in Sect. 5.2. Then, we apply each reduction technique to reduce every dataset to 3, 50, and 150 dimensions. For t-SNE, we computed only 3-dimensional reduction because it was too expensive to compute the others.

5.4.3 Results

Table 3 shows the estimated W_2^2 to check whether it would still capture the increasing bias after the reduction. Furthermore, we tested whether the image ranking order of the Kantorovich potentials is maintained compared to the ranking induced by the original features by computing the Spearman correlation between the two types of ranking.

Apart from SE, dimensionality reduction techniques successfully capture the selection bias in data. Nonetheless, the results tend to be different between techniques; Isomap and t-SNE tend to exacerbate bias while PCA reduces it slightly. Similarly, apart from SE, dimensionality reduction tends to preserve the rankings quite well, as shown by the high Spearman correlation before and after reduction. Isomap has the best score, especially for highly biased data (30% or more of the people wearing hats in the biased sample). But, it is PCA that best preserves the order of Kantorovich potentials.

6 Discussion

Both the considerations in Sect. 5.1 and the theoretical guarantees in Sects. 3.2 and 3.3 present our method as a valid alternative to existing baselines. For the model-based baseline (Torralba and Efros 2011), we only showed that there exist cases in which it fails and not that we always outperform it. Nonetheless, the underlying principles of OT theory ensure that our method intrinsically overcomes issues that would arise from heuristics. We also showed an equivalence to the image ranks of Kwegyir-Aggrey et al. (2021), but the use of W_2^2 is a better global bias measure compared to more naive aggregation of individual sample bias.

Overall, we experimentally demonstrated that W_2^2 can capture (selection) data biases quite well. The distribution and corresponding order of the Kantorovich potentials adequately represents image bias at least for the CelebA splits. Biased MNIST revealed itself as a more difficult dataset to analyse using the Kantorovich potentials, this might be due to strong correlations between the different attributes that causes the biases for the various digits but needs further exploration. Finally, appropriate dimensionality reduction methods do not significantly affect the ranks order of Kantorovich potentials and -ultimately- the applicability of our pipeline.

Before closing this work, we point out that the experiments in Sects. 5.2 and 5.3, as well as the case study in Sect. 4 yield purely observational results, which is in part inevitable for a methodology that aims to assist qualitative data analysis. It is indeed not possible to assess whether chosen attributes (e.g., Waring_Hat, Wearing_Necktie, Eyeglasses, and Smiling for CelebA) are the *causes* of the measured bias; other causes could be correlated to such attributes and may not be immediately apparent to visual inspection.

We also remark that the user study in Sect. 4.5 presents some weaknesses. Indeed, it is performed on a very small set of participants from a small number of research groups. Nonetheless, we believe that it gives a clear idea of how our work can support bias detection and on the extent to which it is subject to confirmation bias.

Finally, for the sake of computational tractability, we employ a non-exact method to compute W_2^2 and the Kantorovich potentials. Nonetheless, the approach we employ can be considered as state-of-the-art (Korotin et al. 2021). During this approximation, the efficacy of ICNNs in Sect. 3.1 depends on the hyperparameter choices (e.g., the optimiser) and therefore these need to be tuned anew for each dataset being explored.

7 Conclusions

In this work, we introduced a framework for bias detection based on OT theory. Its strong theoretical guarantees and ability to be incorporated into various data mining pipelines makes it a valid option for bias detection. We also conducted a case study that others can replicate to integrate our proposed approach into bias-aware data mining pipelines to derive hypothesis on possible biases similar to the ones we reveal for the CelebA dataset.

Analysis that uses our framework would not be final but rather serve as a *prima facie* evidence of the presence of bias in the data. Further investigation should either confirm or disprove our method's findings. There are few methods in the literature that help make such hypotheses, and we try to fill this gap.

Future work can be devoted to the refinement of our method, for example through further development of accurate OT approximation algorithms. Finally, our framework could be integrated in other bias discovery pipelines, and motivate new preprocessing approaches for bias mitigation.

Appendix A Mathematical background

In this appendix, we provide the reader with mathematical background that helps gain a more in-depth understanding of Sects. 2.2, 3.4, and 3.2.

A.1 Optimal transport details

We provide some standard definitions of metric space and measure theory, and the reader can refer to Salamon (2016) and Villani (2008) for further details.

Definition 2 (Metric space) Given a set X, a function $d : X \times X \to \mathbb{R}_{\geq 0}$ is called a *distance* or a *metric* if the following conditions hold:

- $d(x, y) = 0$ iff $x = y \quad \forall x, y \in X$
- $d(x, y) = d(y, x) \quad \forall x, y \in X$
- $d(x, y) \leq d(x, z) + d(z, y) \quad \forall x, y, z \in X$

The couple (X, d) is called a *metric space*.

Definition 3 (Measurable space) Given a set X, a collection $\mathcal{A}$ of subsets of X is called **σ-algebra** if:

1. $X \in \mathcal{A}$
2. if $A \in \mathcal{A}$ then $X \setminus A \in \mathcal{A}$
3. Every countable union of subsets in $\mathcal{A}$ is in $\mathcal{A}$

The couple $(X, \mathcal{A})$ is called a *measurable space*.

Definition 4 (Measurable function Salamon 2016) A function $f : (X, \mathcal{A}) \to (Y, \mathcal{B})$ between measurable spaces is said to be **measurable** if for every $B \in \mathcal{B}$ then $f^{-1}(B) \in \mathcal{A}$.

Definition 5 (Measure) Given a measurable space $(X, \mathcal{A}$, we call **measure** a function

$$\mu : \mathcal{A} \to [0, +\infty]$$

such that

1. μ is σ-additive, i.e., given a a sequence of pairwise disjoint union of sets in $\mathcal{A}$ then

$$\mu(\bigcup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mu(A_i)$$

2. There exist a measurable set $A \in \mathcal{A}$ such that $\mu(A) < \infty$

The triple $(X, \mathcal{A}, \mu)$ is called a *measure space*.

Definition 6 (Push-forward measure) Given a measure space $(X, \mathcal{A}, \mu)$ and a map $f : X \to Y$, then we define the *push-forward measure* $f_{\#}\mu : f_{\#}\mathcal{A} \to [0, +\infty]$ such that

$$f_{\#}\mu(B) = \mu(f^{-1}(B))$$

where $f_{\#}\mathcal{A}$ is the σ-algebra $\{B \subseteq Y | f^{-1}(B) \in \mathcal{A}\}$. Note that when Y is endowed with the σ-algebra $f_{\#}\mathcal{A}, f$ is automatically a measurable function.

Definition 7 (Coupling of two measures) Given two measure spaces $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \nu)$, we say that a measure π on the product space $X \times Y$ is a **coupling** if $(p_X)_{\#}\pi = \mu$ and $(p_Y)_{\#}\pi = \nu$ where p_X and p_Y are the standard projections onto X and Y, respectively. We, alternatively, say that *pi* has **marginals** μ and ν. We indicate the set of couplings for μ and ν as $\Pi(\mu, \nu)$.

Definition 8 (Polish Space) A metric space (X, d_X) is called **Polish** if it is complete and separable with respect to the topology induced by the d_X.

A.2 Mapper details

To gain a better understanding of Mapper, we revise some basic notions of topology (Munkres 2000).

Definition 9 Given a set X and a collection of subsets $T = \{U_i \subseteq X\}$ such that:

1. $\emptyset \in T$
2. Any arbitrary union of elements of T is in T
3. Any finite intersection of elements of T is in T

The subsets in T are called *open sets*, T is called a *topology* on X, and X is called a *topological space*.

The above definition enables the development of the entire theory of topology and the study in a very precise way of the shape of spaces. In particular, it enables the definition of what constitutes a *continuous function*. Intuitively, the latter is a function that maps a space into another without "ripping" it.

Definition 10 A function $f : (X, T_X) \to (Y, T_Y)$ between topological spaces is called *continuous* if the *pre-image* $f^{-1}(U) := \{x \in X | f(x) \in U\}$ of $< n$ open set $U \in T_Y$ is in turn an open set in T_X.

Definition 11 A set of open sets $\mathcal{U} = \{U_i\} \subseteq T_Y$ is called an *open covering* of Y if $\bigcup U_i = Y$. Note that the pre-image of the sets in $\mathcal{U}$ along a continuous function $f : (X, T_X) \to (Y, T_Y)$ is a covering of X. Given a covering $\mathcal{U}$, we can construct a graph $N(\mathcal{U})$ called the *nerve complex* in the following way: each open $U \in \mathcal{U}$ corresponds to a node and two nodes are linked by an edge if and only if the intersection between the respective open sets is non-empty.

Finally, a topological space (X, T) is called *disconnected* if there exist at least two non-empty open sets $V, U \in T$ such that $X = V \cup U$ and $V \cap U = \emptyset$. In such a case, U and V are called *connected components* of the space X.

Appendix B Implementation details

In this appendix, we provide the reader with the details about the implementation of the Kantorovich potential approximation.

B.1 Case study

For each of the three feature spaces, we trained two 4-layers ICNN ϕ and ψ with 512 neurons per layer and input dimension equal to 3. Both the networks have Leaky ReLU activation functions. They adopt the regularised objective presented by Makkuva et al. (2020) and use a lambda of 0.5. Both networks are trained over 25 epochs using an RMSProp optimiser with hyperparameters $\alpha = 0.99$, momentum 0.5, and learning rate 0.0001 which is further halved after 20 epochs.

The network ψ was trained 5 times per iteration of ϕ. While ϕ is fed with batches sampled directly from a set of data points X (i.e., we consider $\mu = \sum_{x_i \in X} \delta_{x_i}$), Brenier's theorem (described in Sect. 2.2) requires the distribution ν to be absolutely continuous. Hence, we sample the batches from a sum of Gaussians centred in the points of Y. In practice, we sample a batch $(x_1, ..., x_n)$ from Y and for every x_i we re-sample a point from $\mathcal{N}(x_i, \Sigma_i)$. In our experiments, $\Sigma_i = 0.001 \cdot I_3$ for every i, where I_3 is the 3×3 identity matrix.

B.2 Experiments

For the baseline comparison experiments on both synthetic datasets, the two 4-layers ICNNs ϕ and ψ were trained with similar settings as in the case study. In particular, they were trained for 25 epochs using an RMSProp optimiser with learning rate of 0.0001 which is further cut by after 20 epochs. Again, ψ was trained 5 times per iteration of ϕ. Since samples were already obtained from continuous distributions, this time we did not re-sample Y from the sum of Gaussians.

During experimentation to assert the ability of W_2^2 and Kantorovich potentials in capturing global and data sample bias respectively, we used the same exact configuration as in the case study. The only difference was that the input feature dimensions were 512.

Dimensionality reduction experiments, also use the same configuration as in the case study, apart for the input dimension. This time, RMSProp did not converge for the spaces reduced with SE, and in those cases we used Stochastic Gradient Descent with the same learning rate and momentum.

Appendix C Additional mapper analysis

Here we show the results of the Mapper analysis for the Autoencoder and FaceNet feature presented in Sect. 4 (Figs. 8, 9).

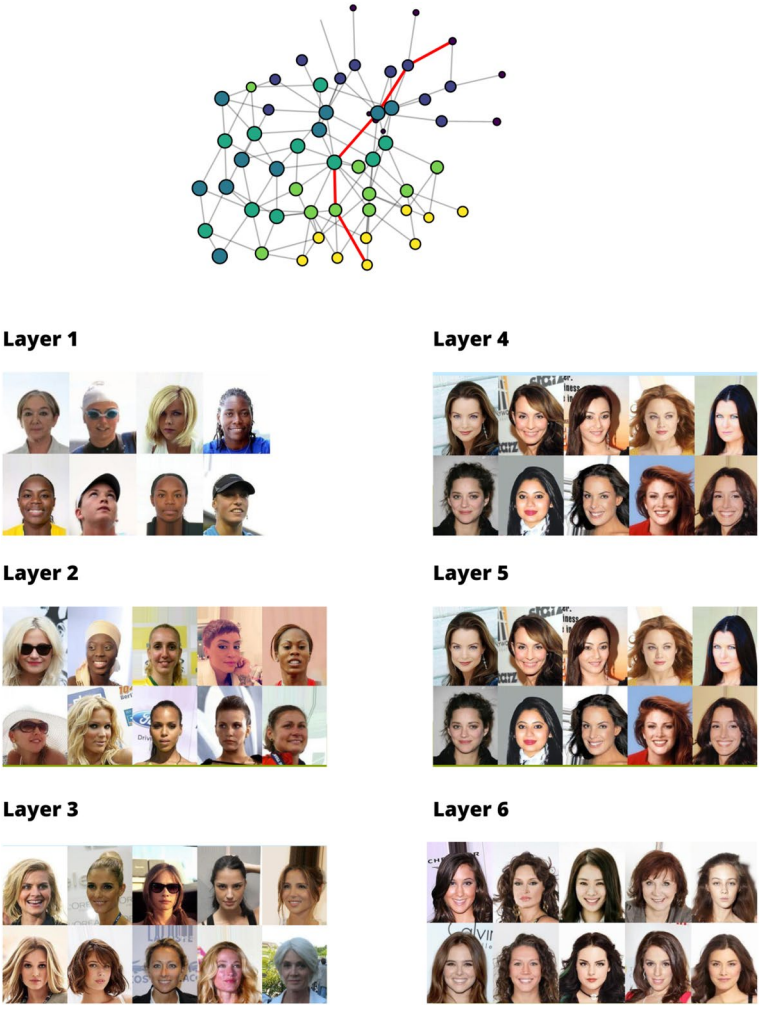


Fig. 8 A path in the nerve complex constructed by Mapper for the three-dimensional Autoencoder features

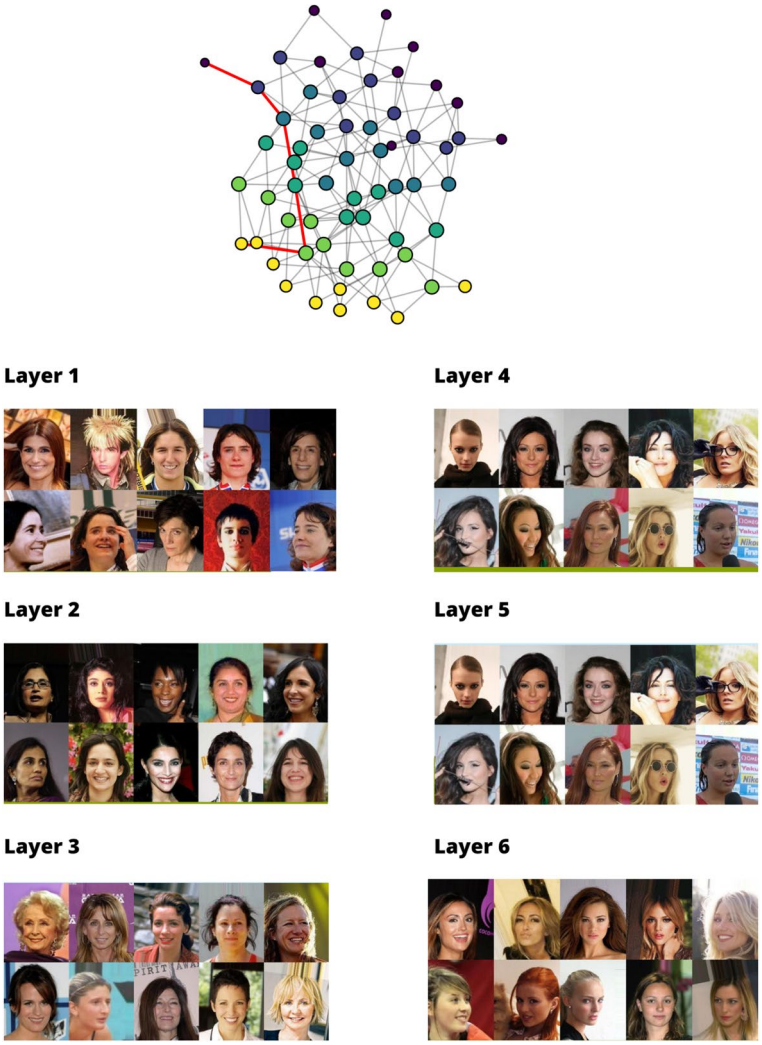


Fig. 9 A path in the nerve complex constructed by Mapper for the three-dimensional FaceNet features

Appendix D User study

To corroborate our observations in Sect. 4.4 concerning the ability of our approach to produce helpful visualizations, we conducted a case study. In this, 30 participants were gathered on a voluntary basis from the research groups of the authors. We divided them into experiment and control groups. Participants joined the study at different times -essentially based on their research group- and we stratified the separation between batches with a clustered randomisation procedure; every time a group of users joined the study, we sampled half of them without replacement and assigned them to the experiment group and left the rest to the control. Of the 30

Table 4 Answers to the quantitative questions in the survey

Ranking	Q1	Q3	Q5	Q7 (mean)	Q7 (std)
ResNet18	75% Yes	58.3% Yes	91.7 % Yes	5.75	1.16
Autoencoder	91.7 % Yes	66.7 Yes	91.7 % Yes	5.92	1.32
FaceNet	91.7 % Yes	58.3 % Yes	91.7 % Yes	6.0	1.0
Random control 1	76.9 % Yes	69.2 % Yes	53.8 % Yes	4.15	1.61
Random control 2	84.6 % Yes	53.8 % Yes	53.8 % Yes	4.31	1.94
Random control 3	61.5 % Yes	69.2 % Yes	69.2 % Yes	3.62	1.86

participants, only 25 completed the survey, of which 12 were assigned to the experiment group and 13 to the control group.

The experiment group was shown a total of six batches with 40 images each sampled from the top and bottom 1% of images ranked according to the Kantorovich potentials for ResNet, Autoencoder and FaceNet features (Reduced to dimension 3 with PCA) of the CelebA dataset. The control group was shown the same number of images and with the same visual organisation, though uniformly sampled from the whole dataset.

Irrespective of which group they belong to, each participant ended up being asked to answer the following questions for three different image batches:

Q1) Look at the first batch of 40 images (sampled from the top 1% for the first group, sampled uniformly for the other group). Can you find any patterns in the images?
Q2) If so, describe the patterns you detected.
Q3) Look at the second batch of 40 images (sampled from the bottom 1% for the first group, sampled uniformly for the other group). Can you find any patterns in the images?
Q4) If so, describe the patterns you detected.
Q5) Compare the two batches, does one of the two show a more biased representation of the subjects?
Q6) If so, elaborate.
Q7) Rate from 1 (absent) to 7 (extremely high) the amount of bias found.

Table 4 summarises the answers to quantitative questions (Q1, Q3, Q5, and Q7). To begin with, all participants were more prone to spot bias in the first batch (with the exception of the third pair of batches for the control group). Nonetheless, the experiment group was more likely to respond that one of the two batches had a more biased representation of women. By extension, for (Q7) the experiment group responded with higher score on average and lower standard deviation, where the latter indicates a general consensus among participants.

Table 5 summarises the answers to qualitative questions (Q2 and Q4). We report the patterns identified by more than one participant sorted by the number of participants that detected such pattern. The control group always reported

Table 5 Answers to (Q2) and (Q3) sorted by the number of participants that indentified respective patterns. We report only patterns identified by at least two participants

Batch	Patterns
Top 1% ResNet18	long-hair, white people, no black people, young, attractive, filtered, make up, others
Bottom 1% ResNet18	short or tied hair, head gears, people from minorities, glasses, athlets, others
Top 1% Autoeconder	dark hair, long hair, attractive, young, frontal pose, white people, make up, others
Bottom 1% Autoencoder	lighter hair, others
Top 1% FaceNet	attractive, make-up, young, frontal pose, others
Bottom 1% FaceNet	older, possible mislabbelling, others
Random control 1.1	make-up, white people, smiling, young, long-hair, no black people, attractive, others
Random control 1.2	white people, smiling, young, make-up, long-hair, no black people, others
Random control 2.1	white people, make-up, young, long-hair, dark-hair, smiling, others
Random control 2.2	white people, young, make-up, long-hair, smiling, others
Random control 3.1	white people, smiling, young, make-up, long-hair, others
Random control 3.2	white people, smiling, blonde, make-up, long-hair, young, others

Table 6 Summary of the answers to (Q6). We report only patterns identified by at least two participants

Batches	Patterns
ResNet18	The 2nd batch shows more diversity with respect to skin colour and age, 1st batch shows more (conventionally) attractive women, others
Autoeconder	The 2nd batch shows more diversity in terms of skin colour, facial characteristics, profession, hairstyle, and pose, others
FaceNet	The 2nd batch is more diverse with respect to skin colour and age, others
Random control 1.1 vs 1.2	The 1st batch shows only white women, others
Random control 2.1 vs 2.2	The 1st batch shows more conventionally attractive women, the 1st batch shows more (conventionally attractive) women, others
Random control 3.1 vs. 3.2	1st batch is more diverse, others

the same patterns, which are very prominent in CelebA and extremely common among its images labelled as female. Whereas the experiment group detected a variety of patterns that are very different between images with high and low potentials. Table 6 summarises the answer to (Q6). Similarly to (Q7), the experiment group was more likely to identify biases between the two batches. Nonetheless, different patterns among the random samples were occasionally identified by people in the control group, too. Therefore, we cannot entirely exclude the possibility to fall into confirmation bias when analysing a dataset with our method.

Appendix E Additional experiment results

Here we show the results of some additional experiments Fig. 10.

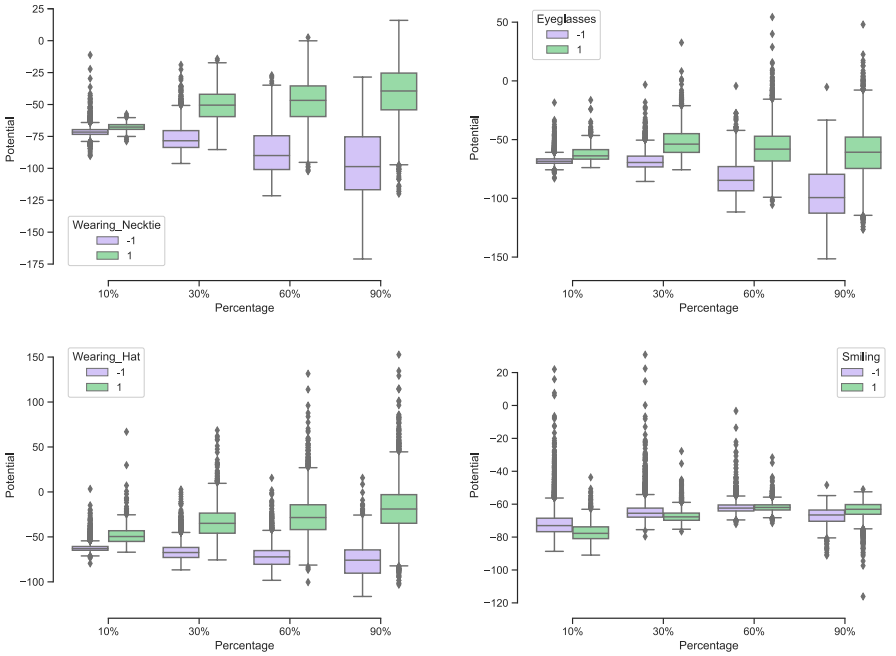


Fig. 10 Box plots of the value of the Kantorovich potential with respect to the attributes Wearing_Necktie, Eyeglasses, Wearing_Hat, and Smiling

Acknowledgements We would like to thank Antonio Ferrara and Carlos Mougan for their feedback and fruitful discussions. This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS - Artificial Intelligence without Bias" and under grant agreement number 101070285 for the project "MAMMOth - Multi-Attribute, Multimodal Bias Mitigation in AI Systems". This work reflects only the authors' views and the European Research Executive Agency (REA) is not responsible for any use that may be made of the information it contains.

Funding Open access funding provided by HEAL-Link Greece.

References

Amos B, Xu L, Kolter JZ (2016) Input convex neural networks. CoRR arxiv:1609.07152

Balakrishnan G, Xiong Y, Xia W, Perona P (2020) Towards causal benchmarking of bias in face analysis algorithms. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer vision - ECCV 2020–16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XVIII, vol 12363. Springer, Lecture notes in computer science, pp 547–563

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning: limitations and opportunities. fairmlbook.org. http://www.fairmlbook.org

Belkin M, Niyogi P (2003) Laplacian eigenmaps for dimensionality reduction and data representation. Neural comput 15(6):1373–1396. https://doi.org/10.1162/089976603321780317

Berendt B, Preibusch S (2014) Better decision support through exploratory discrimination-aware data mining: foundations and empirical evidence. Artif Intell Law 22(2):175–209. https://doi.org/10.1007/s10506-013-9152-0

Buolamwini J, Gebru T (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler SA, Wilson C (Eds.), Conference on fairness, accountability and transparency, FAT 2018, 23-24 February 2018, New York, Volume 81 of proceedings of machine learning research, pp. 77–91. PMLR

Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp. 67–74. IEEE Press

Chiappa S, Jiang R, Stepleton T, Pacchiano A, Jiang H, Aslanides J (2020) A general approach to fairness with optimal transport. In The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, February 7-12, 2020, pp. 3633–3640. AAAI Press

Chiappa S, Pacchiano A (2021) Fairness with continuous optimal transport. CoRR arxiv:2101.02084

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1023/A:1022627411411

Cui M et al (2020) Introduction to the k-means clustering algorithm based on the elbow method. Account Audit Financ 1(1):5–8

Cuturi M (2013) Sinkhorn distances: lightspeed computation of optimal transport. In: Burges C, Bottou L, Welling M, Ghahramani Z, Weinberger K (eds) Advances in neural information processing systems, vol 26. Curran Associates Inc

Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In 2009 IEEE computer society conference on computer vision and pattern recognition (CVPR 2009), 20-25 June 2009, Miami, pp. 248–255. IEEE computer society

Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring adult: new datasets for fair machine learning. Adv Neural Inf Process Syst 34:6478–6490

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226

Fabbrizzi S, Papadopoulos S, Ntoutsi E, Kompatsiaris I (2022) A survey on bias in visual datasets. Comput Vis Image Underst 223:103552. https://doi.org/10.1016/j.cviu.2022.103552

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach HM, III HD, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92. https://doi.org/10.1145/3458723

Gordaliza P, Barrio ED, Fabrice G, Loubes JM (2019), 09–15 Jun. Obtaining fairness using optimal transport theory. In K. Chaudhuri and R. Salakhutdinov (Eds.), Proceedings of the 36th international conference on machine learning, Volume 97 of proceedings of machine learning research, pp. 2357–2365. PMLR

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, June 27-30, 2016, pp. 770–778. IEEE computer society

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:498–520

Jiang R, Pacchiano A, Stepleton T, Jiang H, Chiappa S (2019) Wasserstein fair classification. In A. Globerson and R. Silva (Eds.), Proceedings of the thirty-fifth conference on uncertainty in

artificial intelligence, UAI 2019, Tel Aviv, July 22-25, 2019, Volume 115 of proceedings of machine learning research, pp. 862–872. AUAI Press
Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33
Kamiran F, Calders T, Pechenizkiy M (2010) Discrimination aware decision tree learning. In 2010 IEEE international conference on data mining, pp. 869–874. IEEE
Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2012, Bristol, September 24-28, 2012. Proceedings, Part II 23, pp. 35–50. Springer
Kärkkäinen K, Joo J (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
Korotin A, Li L, Genevay A, Solomon JM, Filippov A, Burnaev E (2021) Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. In: Ranzato M, Beygelzimer A, Dauphin Y, Liang P, Vaughan JW (eds) Advances in neural information processing systems, vol 34. Curran Associates Inc, pp 14593–14605
Kwegyir-Aggrey K, Santorella R, Brown SM (2021) Everything is relative: Understanding fairness with optimal transport. CoRR arXiv:2102.10349
Li FF, Andreetto M, Ranzato M, Perona P (2022) Caltech 101
Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In Proceedings of international conference on computer vision (ICCV)
Makkuva AV, Taghvaei A, Oh S, Lee JD (2020) Optimal transport mapping via input convex neural networks. In Proceedings of the 37th international conference on machine learning, ICML 2020, 13-18 July 2020, Virtual event, Volume 119 of proceedings of machine learning research, pp. 6672–6681. PMLR
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):3457607. https://doi.org/10.1145/3457607
Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in Faces. arXiv:1901.10436 [cs.CV]
Miroshnikov A, Kotsiopoulos K, Franks R, Kannan AR (2022) Wasserstein-based fairness interpretability framework for machine learning models. Mach Learn 111(9):3307–3357. https://doi.org/10.1007/s10994-022-06213-9
Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Ann Rev Stat Appl 8(1):141–163. https://doi.org/10.1146/annurev-statistics-042720-125902
Munkres JR (2000) Topology. Prentice Hall Inc, New Jersy
Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E, Kompatsiaris I, Kinder-Kurlanda K, Wagner C, Karimi F, Fernandez M, Alani H, Berendt B, Kruegel T, Heinze C, Broelemann K, Kasneci G, Tiropanis T, Staab S (2020) Bias in data-driven artificial intelligence systems–an introductory survey. WIREs Data Min Knowl Discov 10(3):e1356. https://doi.org/10.1002/widm.1356
Panda R, Zhang J, Li H, Lee JY, Lu X, Roy-Chowdhury AK (2018) Contemplating visual emotions: understanding and overcoming dataset bias. In ECCV
Ribeiro MT, Singh S, Guestrin C (2016) "why should I trust you?": Explaining the predictions of any classifier. In Krishnapuram B, Shah M, Smola AJ, Aggarwal CC, Shen D,. Rastogi R (Eds.), Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13-17, 2016, pp. 1135–1144. ACM
Salamon DA (2016) Measure and Integration (1 ed.). EMS Press, Berlin
Sattigeri P, Hoffman SC, Chenthamarakshan V, Varshney KR (2019) Fairness gan: Generating datasets with fairness properties using a generative adversarial network. IBM J Res Dev 63(4/5):1–3
Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. CoRR arXiv:1503.03832
Shrestha R, Kafle K, Kanan C (2022) Occamnets: mitigating dataset bias by favoring simpler hypotheses
Singh G, Memoli F, Carlsson G (2007) Topological methods for the analysis of high dimensional data sets and 3D object recognition. In: Botsch M, Pajarola R, Chen B, Zwicker M (eds) Eurographics symposium on point-based graphics. The Eurographics Association
Steed R, Caliskan A (2021) Image representations learned with unsupervised pre-training contain human-like biases. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, New York, pp. 701-713. Association for computing machinery

Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In proceedings of the 27th international conference on neural information processing systems - Volume 2, NIPS'14, Cambridge, MA, pp. 1988–1996. MIT Press

Tenenbaum JB, de Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–23

Tommasi T, Patricia N, Caputo B, Tuytelaars T (2015) A deeper look at dataset bias. In: Gall J, Gehler PV, Leibe B (eds) Pattern recognition - 37th German conference, GCPR 2015, aachen, Germany, October 7–10, 2015, proceedings, vol 9358. Lecture notes in computer science. Springer, pp 504–516

Torralba A, Efros AA (2011) Unbiased look at dataset bias. In The 24th IEEE conference on computer vision and pattern recognition, CVPR 2011, Colorado Springs, CO, 20-25 June 2011, pp. 1521–1528. IEEE computer society

Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605

Villani C (2008) Optimal transport: Old and new

Villani C (2003) Topics in optimal transportation. Graduate studies in mathematics. American Mathematical Society, Providence

Wang A, Liu A, Zhang R, Kleiman A, Kim L, Zhao D, Shirai I, Narayanan A, Russakovsky O (2022) REVISE: a tool for measuring and mitigating bias in visual datasets. Int J Comput Vis 130(7):1790–1810. https://doi.org/10.1007/s11263-022-01625-5

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th international conference on world wide web, pp. 1171–1180

Zehlike M, Hacker P, Wiedemann E (2020) Matching code and law: achieving algorithmic fairness with optimal transport. Data Min Knowl Discov 34(1):163–200. https://doi.org/10.1007/s10618-019-00658-8

Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (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, Copenhagen, Denmark, pp. 2979–2989. Association for computational linguistics