The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R^2HCAI)

A Case Study in Fairness Evaluation: Current Limitations and Challenges for Human Pose Estimation

Julienne LaChance* **William Thong*** **Shruti Nagpal** **Alice Xiang**

Sony AI

julienne.lachance@sony.com, william.thong@sony.com, shruti.nagpal@sony.com, alice.xiang@sony.com

Abstract

With a growing interest in understanding the risks of machine learning models, fairness evaluations help to assess potential discrimination. For example, fairness evaluations can reveal differences in a model's performance across different genders, age groups or skin tones. However, practical limitations and challenges inhibit such assessments, especially in the context of human-centric tasks in computer vision. In this paper, we focus on human pose estimation and pinpoint the difficulties of performing a fairness evaluation with currently available datasets. We first highlight the lack of demographics labels in the current literature, which prevents practitioners from easily evaluating model biases. Second, based on this observation, we annotate the validation set of COCO-Keypoints for demographic labels and reveal the inherent data bias for pose estimation. Third, we evaluate several pose estimation models on our annotated subsets to gain insights on the challenges pertaining the operationalization of fairness evaluations. We finally discuss recommendations in the fairness space to overcome the identified barriers to utility.

1 Introduction

With an increasing societal awareness on the risks of machine learning models (Andrus et al. 2021), fairness evaluations become essential to assess and document any potential discrimination against specific protected groups. This is particularly relevant for human-centric tasks, where the performance of computer vision models can differ across different genders, age groups, or skin tones (Buolamwini and Gebru 2018; Raji and Buolamwini 2019; Wilson, Hoffman, and Morgenstern 2019). For example, Buolamwini and Gebru (2018) crucially identify that images of darker-skinned females have a higher error rate in commercial computer vision systems. Likewise, Wilson, Hoffman, and Morgenstern (2019) point out the lower predictive performance of person detectors in images of darker-skinned individuals. Yet, there exist limitations and challenges to perform an appropriate and relevant fairness evaluation for human-centric tasks.

To illustrate the practical difficulties in performing a fairness evaluation, we focus on 2D human pose estimation as a case study due to its utility in a wide range of higher-level human-centric visual tasks. Pose estimation can serve as a basis to recognize human actions (*e.g.*, (Zhang et al. 2019a)), perform person re-identification (*e.g.*, (Su et al. 2017)), or detect human-object interactions (*e.g.*, (Yao and Fei-Fei 2010)). As the application space utilizing pose estimation models expands, the potential risks of biased pose models increases correspondingly; in the healthcare space, for instance, hospital fall-detection algorithms may be less able to identify injured patients who are female or darker-skinned, resulting in delayed care. Nevertheless, fairness evaluators in the space of human pose estimation face several limitations. Despite the widespread use of the 2D human pose estimation task, datasets usually do not include any comprehensive demographic labels, either self-reported by the subjects or labeled by how annotators perceive them. Such absence prevents reseachers and practitioners from performing any fairness evaluation to understand potential model biases, which in the end affects the end users.

Consider, for example, the Common Objects in Context (COCO) dataset (Lin et al. 2014), which has led to tremendous progress in human pose estimation (He et al. 2017; Li et al. 2019; Liu and Mei 2022), but also image captioning (Vaswani et al. 2017; Huang et al. 2019; Li et al. 2020; Hossain et al. 2019) as well as object detection and segmentation (Liu et al. 2020; Minaee et al. 2021). Even though COCO has been widely adopted, it presents several limitations for fairness evaluation utilization. To observe this, we select relevant images in the validation set for single pose estimation, and semi-manually annotate the resulting images for demographic labels, resulting in the COCO-Keypoints-Demographics (COCO-KD) subset. Statistics then show an over-representation of males and lighter-skinned individuals; and an under-representation of females, darker-skinned and older individuals. Such inherent biases in COCO question its relevance for a fairness evaluation.

The limitations regarding demographic labels or imbalanced datasets challenge the operationalization of fairness evaluations. Indeed, transparency documents, such as datasheets (Gebru et al. 2021) or model cards (Mitchell et al. 2019), would require the identification of subpopulations to report how representative a dataset is or whether a model exhibit performance discrepancies. To understand the challenges involved with such limitations, we evaluate several 2D human pose estimation models, widely used in academia or industry on the proposed COCO-KD subset. While a straightforward evaluation highlights potential model biases

*These authors contributed equally.

towards gender, age or skin tone, we also question the reliability of these conclusions by exploring how model biases can be influenced by changing the evaluation set.

The main contribution of the paper is the identification of current limitations and challenges in operationalizing fairness evaluations for a human-centric vision task, namely the 2D human pose estimation task. First, we describe the lack of demographic labels in publicly available human pose datasets. This lack of demographic labels restricts the ability to perform a fairness analysis of datasets and models. Second, we address current issues by providing fairness annotations on a subset of the validation set of COCO, which will be made public. This enables an assessment of existing biases in human pose estimation models. Still, using COCO as a dataset for fairness evaluation has its limits, as the dataset is highly imbalanced. Third, we evaluate several models for 2D human pose estimation. Leveraging our fairness annotations, we identify potential model biases present in all evaluated models. We further explore the reliability and meaningfulness of these fairness evaluations as different subsets of the dataset could influence the fairness conclusions. Finally, we provide future recommendations towards a better operationalization of fairness evaluations in the context of human-centric vision tasks, with an example from human pose estimation.

The rest of the paper is structured as follows: Section 2 reviews related works in fairness evaluation and pose estimation; Section 3 highlights the lack of demographic labels in the literature; Section 4 presents the data bias in COCO-KD after our semi-manual annotations; Section 5 reports the fairness evaluations of pose estimation models while Section 6 shows how these fairness evaluations could be influenced to change the bias direction; and Section 7 provides recommendations and concluding remarks.

2 Related Work

Advancements in the field of artificial intelligence have led to their widespread adoption. Their fairness evaluation remains a challenge (Fabbrizzi et al. 2022). Researchers have highlighted this challenge in various domains ranging from facial analysis (Buolamwini and Gebru 2018; Khalil et al. 2020), hiring systems (Parasurama and Sedoc 2021; Raghavan et al. 2020) to automated caption generation models (Zhao, Wang, and Russakovsky 2021). Significant efforts have also been made to understand (Nagpal et al. 2019; Wang, Narayanan, and Russakovsky 2020) as well as mitigate this effect (Tang et al. 2021; Wang et al. 2020; Karkkainen and Joo 2021; Thong and Snoek 2021). However, limited work has been done to discuss and analyze the importance of and challenges to perform and operationalize fairness audits, especially in the human-centric context. We discuss related work for fairness evaluation and audits, and pose estimation.

Fairness evaluation. Mitchell et al. (2019) proposed providing detailed documentation with the models that are released and presented a framework, termed as *model cards*. The model cards are documents for corresponding models which provide benchmark evaluation in various conditions (including different subgroups such as gender and demographic groups) that are relevant to the intended application and use. Raji et al. (2020) introduced a framework for end-to-end auditing of AI systems to be applied throughout the development process. The framework is designed to reduce the *accountability gap* in the deployment of AI systems. Holland et al. (2020) proposed *dataset nutrition labels*, a diagnostic framework for standardized data analysis which provides an exhaustive overview of the dataset used for model development. More recently, similar to the nutrition label and model cards for model, Gebru et al. (2021) proposed datasheets for datasets, a document with information such as motivation, collection process, recommended uses, etc. Datasheets have the same aim to facilitate transparency and accountability. Landers and Behrend (2022) build an interdisciplinary understanding of fairness in AI systems and present psychological audits as a standardized approach to evaluate bias across different categories such as source data, design, development, how information is presented etc. In terms of deploying and operationalizing fairness, Madaio et al. (2020) discuss the importance of how checklists can enable operationalizing fairness of AI models. The authors conducted a co-design process with AI practitioners to design a fairness checklist. They identified concerns associated with checklists and how checklists can be used to provide organizational infrastructure to ad-hoc processes. Additionally, a number of fairness toolkits have been developed to facilitate fairness evalutations, such as Fairness360 (Bellamy et al. 2019) and REVISE (Wang, Narayanan, and Russakovsky 2020). While these fairness analysis strive to provide an objective measure, there exist some subjectivity in how the annotations of the protected attributes are done. Indeed, annotations can sometimes be absent or erroneous (Mehrotra and Celis 2021). In this paper, we build on this literature and highlight potential limitations and challenges associated with operationalizing fairness evaluations.

Pose Estimation is the task of localizing various key points of a human in a given image. It is the base for multiple computer vision tasks such as pose classification, action recognition, sign language classification, human tracking, etc. It has been a widely researched area in computer vision. Munea et al. (2020) present a comprehensive review of 2D human pose estimation. There are broadly two kinds of approaches for pose estimation: (i) top down: a person detector is run to estimate keypoints within the bounding box, and (ii) bottom up: each keypoint is estimated and then all keypoints are connected to form an individual. For example, MoveNet (Chen et al. 2022) follows a bottom-up approach to detect 17 body keypoints. It uses heatmaps to localize human key points and comprises of a feature extractor and prediction model. The feature extractor is built on MobileNetV2 and feature pyramid networks (Lin et al. 2016), while the prediction model is modified CenterNet (Zhou, Wang, and Krähenbühl 2019). PoseNet (Papandreou et al. 2018) is a top-down approach also based on heatmaps which can be used for single or multi-person pose estimation in images and videos. It follows a greedy process to group keypoints when multiple individuals are present in an image.

Dataset	Demographic annotations
COCO (2014)	Augmented with skin tone and gender in (2021); *R.P.O.*
MPII Human Pose (2014)	None
Human3.6M (2014)	Gender
Frames Labeled in Cinema Plus (2014)	None
HumanEVA (2010)	None
DensePose (2018)	Inherits gender and skin tone from COCO; *R.P.O.*
Leeds Sports Pose (2010)	None
JHMDB (2013)	None
CMU Panoptic Studio Dataset (2015)	None
Frames Labeled in Cinema (2013)	None
Unite the People (2017)	None
CrowdPose (2018)	None
PoseTrack (2018)	None
UPenn Action (2013)	None
ITOP Dataset (2016)	None
VGG Human Pose Estimation (2016)	None
OCHuman (2019b)	None
FashionPose (2014)	None
Mannequin RGB and IRS in-bed (2017)	None
UAV Human (2021)	None

Table 1: **Demographic annotations** for pose estimation datasets. Augmented demographic annotations are indicated where known. Annotations to be used for "Research Purposes Only" are indicated via *R.P.O.* Very few large-scale pose datasets have available demographic annotations, and those that do are size limited or usage restricted.

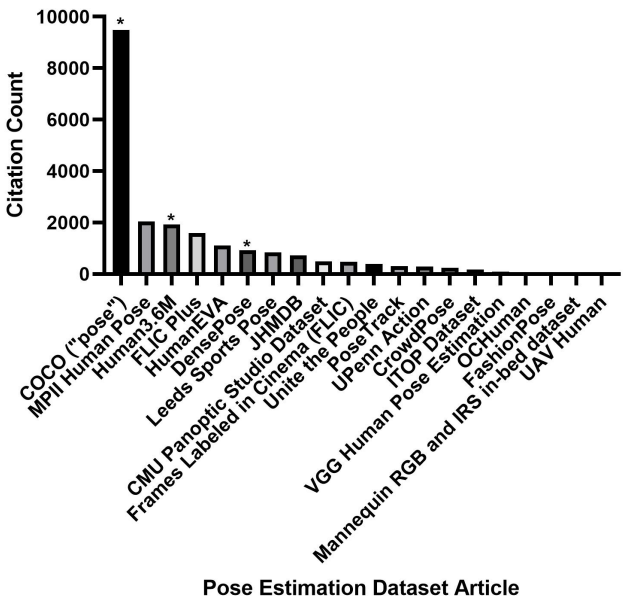


Figure 1: **Citation count** for articles introducing popular pose datasets, per Google Scholar as accessed Aug. 2, 2022. We reduce the COCO citation count by querying only retrieved results with the search term "pose", given that this dataset is utilized for multiple tasks. Asterisks (*) indicate datasets with limited demographic information. The utilization of COCO is shown to be widespread for pose estimation and additional visual tasks.

Cao et al. (2019) proposed OpenPose, a bottom-up deep learning based approach to perform real time 2D pose estimation for multiple individuals. Part affinity fields are used to learn body part associations. OpenPose detects body, foot, hand and facial keypoints. In this paper, we assess the fairness of these models for 2D pose estimation.

3 Lack of Demographic Annotations

Few machine learning datasets come with demographic annotations (Fabris et al. 2022; Madaio et al. 2022). Even in the seminal Gender Shades paper, in which a lower performance of commercial facial classification models occurs on darker-skinned females, Buolamwini and Gebru (2018) ultimately had to manually annotate a novel dataset after encountering limitations with existing benchmark datasets. In this section, we focus on human pose estimation datasets and highlight how they also suffer from a lack of demographic annotations, which are required to perform a comprehensive fairness evaluation. We review the literature and report the publicly-available demographic annotations associated with each member of a set of commonly used human pose estimation datasets.

Method

We select a set of well-known human pose estimation datasets spanning diverse task domains, such as 2D human pose (Lin et al. 2014; Andriluka et al. 2014), 3D human pose (Joo et al. 2015), and dense human pose (Güler, Neverova, and Kokkinos 2018). For each, we report the paper citation count for the article in which the dataset was first introduced as determined via Google Scholar[1], as accessed Aug. 2, 2022. The citation count provides a rough proxy for the relative influence of each dataset. We additionally report the human-centric demographic annotations associated with the original dataset (*e.g.*, age, skin tone, race, ethnicity, age, sex, gender) as well as any known demographic attribute augmentations of the dataset.

Results and Discussion

The list of considered human pose datasets are provided in Table 1, as well as their corresponding article citation counts in Figure 1. We find that the original COCO paper was referenced 25,000+ times, and when restricting the query to the term "pose" this results in 9,490 citations, at least five times more than any other dataset for human pose estimation. COCO is then the most widely used dataset for pose and beyond (Xu, Tasaka, and Yamaguchi 2021). For comparison, the ImageNet image classification database (Deng et al. 2009) and challenge (Russakovsky et al. 2015) papers were cited 41,274 and 31,626 times, respectively, while CIFAR-10 (Krizhevsky, Hinton et al. 2009) was referenced 16,574 times.

[1]https://scholar.google.com

We determine that demographic annotations (*e.g.*, gender, age, skin tone) are rare for human pose datasets, even when those datasets are highly influential within the computer vision community. The original COCO, like most datasets, has no associated demographic annotations. While gender annotations are available for the Human3.6M dataset, the population consists of only 11 actors, resulting in an insufficient sample size for a comprehensive fairness evaluation. Overall, even though demographic annotations might be seldomly available, they are not very comprehensive as they do not include other types of sensitive attributes such as height, weight, pregnancy, disability status, etc.

Given that human pose datasets have not been constructed with the intent of collecting demographic annotations from the start, researchers have proposed to augment COCO and the related DensePose by using Amazon Mechanical Turk workers to add perceived gender and broad skin tone category labels to human subjects (Zhao, Wang, and Russakovsky 2021). However, limitations are still remaining as: (i) these annotations may be unavailable to industry evaluators due to their "Research Purposes Only" status; and (ii) the use of perceived labels introduces an additional source of societal bias and raises questions about the nature and utilization of socially constructed categories (Andrus et al. 2021; Xiang 2022; Hanna et al. 2020).

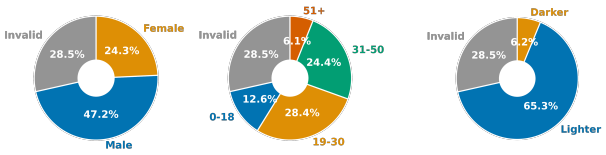
Takeaways. In this context, limitations regarding demographic annotations restrict options of fairness evaluators, who may (i) decide not to perform the assessment, (ii) produce their own annotations, either via tedious manual annotations or costly paid annotators such as crowd-sourcing platform workers, or (iii) utilize inappropriate datasets, such as those which are too small to produce statistically meaningful results. Such limitations may lead practitioners to rely on automated computer vision systems to predict demographic annotations without rigorous manual checks. As a result, this may seriously bias annotations, produce unreliable fairness evaluations, and result in annotations being kept private rather than made publicly available.

4 Imbalanced Demographic Labels

In this section, we highlight the limitations of COCO as a dataset for fairness evaluation, due to its imbalanced demographic distribution. To achieve this, we introduce the COCO-Keypoints-Demographics (COCO-KD) dataset, a subset of the COCO 2D human pose estimation validation set. The construction of this dataset stems from the MoveNet model card (Beletti et al. 2022), which privately annotated the same set of images. We semi-manually augment COCO with binary gender, age, and skin tone annotations; and report hurdles in the annotation process. The distribution of the annotations shows a demographic bias in the collected COCO-KD dataset.

Method

We are inspired by the protocol described in the MoveNet model card (Beletti et al. 2022) to build a subset for fairness evaluation of 2D pose estimation from COCO. Similarly, we start with the COCO Keypoint Dataset Validation Set 2017,



(a) Gender (b) Age (c) Skin tone

Figure 2: **COCO-KD demographic** distribution over 919 images. One third of the images are deemed to be invalid because demographic labels cannot be inferred. For the other two thirds, the distributions show an over-representation of males and light-skinned subjects, as well as an under-representation of older subjects.

and select images where there is only a single individual with at least a bounding box with a surface area of 1000 pixel^2. This results in 919 images out of the 5000 available in the whole validation set. Focusing on images with a single person makes the fairness evaluation more controlled as we solely evaluate the capacity of the model to perform the pose estimation task, rather than person detection and pose estimation simultaneously. While the MoveNet model card utilizes annotations for binary gender, age, and skin tone; these are not publicly available and are collected automatically from internal models, as discussed through personal correspondence with the authors.

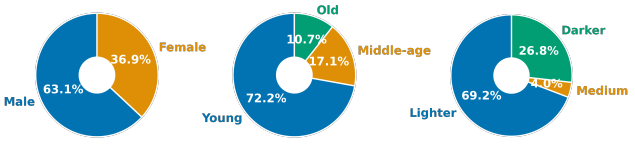
The absence of publicly available demographic labels leads us to collect our own annotations. Similarly, we start with a commercial computer vision system to predict binary gender and age labels, and an internal model measuring the individual typology angle (Chardon, Cretois, and Hourseau 1991) in images to determine the skin tone. For the commercial system, we rely on AWS Rekognition (AWS Rekognition 2022). For all images, the bounding box of the subject is fed as input for demographic label prediction. We discretize the demographic labels as follows: *female* and *male* genders; *0-18*, *19-30*, *31-50* and *51+* age categories; and *lighter* and *darker* skin tones.

To ensure the validity of the annotations, three authors of this paper manually checked the values for every image. (Author 1) and (Author 2) provided perceived gender, age and skin tone annotations independently. (Author 3) resolved conflicts when necessary. All three annotators additionally flagged images where the demographics could not be inferred (*e.g.*, only a hand was visible, or the subject was captured from the back), as well as those which did not contain humans (*e.g.*, a statue). Furthermore, (Author 1) and (Author 3) flagged images with inappropriate content (8 images in total); this notably comprised 4 images with sexualized content and 2 images related to abuse or violence. Our demographic annotations will be made public.[2]

Results and Discussion

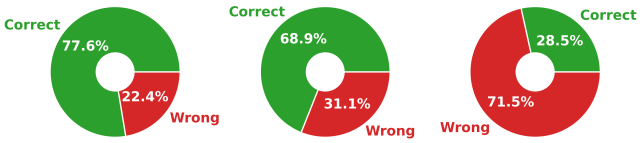
Figure 2 presents the distribution of the demographic labels over the 919 selected images in COCO. We note that

[2]https://github.com/SonyAI/multi_bias_amp



(a) Gender (b) Age (c) Skin tone

Figure 3: **COCO MoveNet demographic** distribution over 919 images, taken from the original model card (Beletti et al. 2022). While these distributions also show imbalanced demographics, the proportions differ from Figure 2, which highlights the importance of manual checks.



(a) Gender (b) Age (c) Skin tone

Figure 4: **COCO-KD annotation correctness** over 657 images after manually verifying the demographic annotations.

657 out of 919 images, or just over two thirds of the initial dataset, are valid, as we only keep images where gender, age, or skin tone labels can be meaningfully inferred. For example, an image which only contains the hand of a person is not enough to determine the age and gender of the subject. We also observe demographic imbalances: male individuals are approximately twice as common as female individuals; older individuals are less frequently represented; and lighter-skinned individuals are present over ten times as frequently as darker-skinned individuals. When looking at intersectional groups, darker-skinned females appear to be present within only 17 of the 657 valid images (2.6%), whereas lighter-skinned males comprise 394 (60.0%).

We observe significant differences between manually-produced annotation results and the automated attribute prediction results. To illustrate these differences, we first compare with the distributions in the MoveNet model card (Beletti et al. 2022) in Figure 3. While ratios are more or less in the same ballpark for gender and age, there is a significant difference for the skin tone distribution. Indeed, our annotations show that COCO-KD has ten times more lighter skins than darker skins, which is much more than the automatic annotations done in COCO MoveNet. Second, we report on how many images we had to manually intervene and change the predicted attribute label during our annotation process in Figure 4. Automated and manual annotations differed by 31.1% for age and 22.4% for gender. Given that the population of darker-skinned female was found to comprise 2.6% of the dataset, the estimated errors from purely automatic attribute prediction could invalidate the results of any fairness evaluation performed with such annotations. Due to these risks, we encourage fairness evaluators to carefully assess annotation quality.

Takeaways. These demographic distributions for the COCO dataset concur with previous literature. Indeed, Zhao, Wang, and Russakovsky (2021) interrogate the demographic imbalance of the COCO 2014 validation set for image captioning. From the crowd-sourced annotations for perceived demographic labels, the dataset appears to be heavily skewed towards male individuals (appearing 2.0x as often as females) and lighter-skinned individuals (appearing 7.5x as often as darker-skinned), with darker-skinned females especially underrepresented. In this paper, we interrogate a subset of the COCO 2017 validation set for pose estimation and therefore expect a similar demographic distribution to the dataset analyzed in (Zhao, Wang, and Russakovsky 2021), as the images are identical across the 2014/2017 versions with differing train/validation/test splits (COCO Tasks 2020). Such demographic imbalances limit the applicability of COCO for fairness evaluation purposes, and particularly for intersectional analysis, since sub-groups across even two attributes (*e.g.*, skin tone, gender) may be vanishingly small (*e.g.*, darker-skinned females only constitute 2.6% of the total images, 17 out of 657 images). It then becomes hard to draw any valid conclusion from these small number of samples. Furthermore, these small sample sizes mean that demographic labels should be reliable and manually validated rather than annotated automatically as this could result in invalid fairness evaluations.

5 Fairness Evaluation

In this section, we evaluate several 2D human pose estimation models on the proposed COCO-KD dataset with demographic labels. We report the overall performance as well as the performance broken down by demographic group (gender, age and skin tone) and their intersectional results.

Method

Setup. We consider the following models: OpenPose (Cao et al. 2019), MoveNet Thunder (Chen et al. 2022) and PoseNet (Papandreou et al. 2018). All models have already been trained, and we simply use them for inference as is. Evaluated models take as input an image and predict the body keypoint location of a single person in the image. All models produce 17 keypoints in 2D for face and body pose estimation. Given that we focus solely on the pose estimation without human detection, we follow the protocol proposed in MoveNet (Chen et al. 2022). We crop the image according the keypoint locations to make sure that the center of the image corresponds to the middle point of the hip area. In case the torso is not visible, we simply resize the image to fit the input size requirements. Once the prediction is done, we map back the results to the original image space to measure the performance of evaluated models.

Metrics. We report the keypoint mean average precision (mAP) and mean average recall (mAR) with object keypoint similarity (OKS) going from 0.50 to 0.95, which is the standard metric used for COCO evaluation and in the pose estimation literature for benchmarking (Lin et al. 2014). Results are computed using the pycocotools toolbox[3] for evaluation.

[3]https://github.com/cocodataset/cocoapi

Images	OpenPose		MoveNet		PoseNet	
	mAP	mAR	mAP	mAR	mAP	mAR
657	79.3	83.2	77.1	80.6	55.9	62.4

Table 2: **2D human pose estimation results** on COCO-KD. OpenPose achieves the highest mAP and mAR, slightly above MoveNet, while PoseNet is far below.

Gender	Images	OpenPose		MoveNet		PoseNet	
		mAP	mAR	mAP	mAR	mAP	mAR
Female	223	78.2	81.7	75.9	79.7	57.1	62.6
Male	434	79.9	83.9	77.9	81.1	55.3	62.3

(a) **Gender**. OpenPose and MoveNet achieve a lower performance for the female group than the male group; while it is the opposite for PoseNet.

Age	Images	OpenPose		MoveNet		PoseNet	
		mAP	mAR	mAP	mAR	mAP	mAR
[0, 18]	116	80.3	83.2	80.4	83.4	59.1	65.0
[19, 30]	261	80.4	84.0	77.4	80.6	51.5	58.0
[31, 50]	224	78.2	82.3	75.3	79.0	59.1	65.0
[51+]	56	78.2	82.9	79.6	81.4	61.4	66.8

(b) **Age**. While discrepencies exist among age groups, they differ depending on the selected models.

Skin tone	Images	OpenPose		MoveNet		PoseNet	
		mAP	mAR	mAP	mAR	mAP	mAR
Lighter-	600	79.3	83.2	77.1	80.6	56.8	63.0
Darker-	57	78.7	82.6	77.4	80.9	47.0	56.0

(c) **Skin tone**. Lighter skins achieve a high performance than darker skins, except for MoveNet where no difference is observed.

Table 3: **Breakdown by demographic labels** on COCO-KD. Models perform differently depending on the demographic sub-group. Such model bias could lead to potential discrimination.

Results and discussion

Table 2 presents the results of 2D human pose estimation models on COCO-KD. Bottom-up approaches such as OpenPose and MoveNet better handle this subset of COCO than top-down approaches such as PoseNet. We observe that OpenPose achieves the highest performance in terms of both mAP and mAR while MoveNet yields slightly lower results. PoseNet has the lowest performance overall, by a large margin, as it struggles to produce predictions with a high OKS score for most images.

Table 3 further breaks down the performance of these models according to the demographic annotations present in COCO-KD. When considering gender in Table 3a, OpenPose and MoveNet tend to achieve a higher score for the male group than the female one, by up to 2 point in mAP or mAR; PoseNet has the reverse effect with the female group achieving a higher mAP. When considering age in Table 3b, performance discrepancies differ among the evaluated models: OpenPose works better for groups below 30 years old while MoveNet prefers under-aged or over-aged groups; and PoseNet tends to struggle for the 19-30 age group. When considering skin tone in Table 3c, OpenPose and PoseNet exhibit a higher performance for lighter skins while MoveNet shows an on-par performance for both groups. Overall, performance discrepancies exist for all the evaluated models regardless of the demographic label being considered. This is an issue as deploying these models at scale could potentially create discrimination towards the under-performing sub-groups.

Skin	Gender	Images	OpenPose		MoveNet		PoseNet	
			mAP	mAR	mAP	mAR	mAP	mAR
Lighter-	Female	206	78.4	81.9	76.2	79.6	58.0	63.0
Darker-	Female	17	78.4	80.0	76.9	81.2	48.3	57.6
Lighter-	Male	394	79.9	83.9	77.9	81.1	56.1	63.0
Darker-	Male	40	79.4	83.7	78.5	80.7	48.0	55.2

Table 4: **Intersectional results** on COCO-KD for gender and skin tone, where males with lighter skin tones tend to always yield a strong performance.

Table 4 presents the results on the intersectional sub-groups. Males with lighter skins tend to also yield a high score regardless of the model being evaluated. For PoseNet, females with lighter skins actually perform better than their male counterparts, which confirms the results present in Table 3. For the other intersectional sub-groups, conclusions are hard to draw given the low number of images. In general, this low number of images makes it difficult to conduct a robust intersectional analysis.

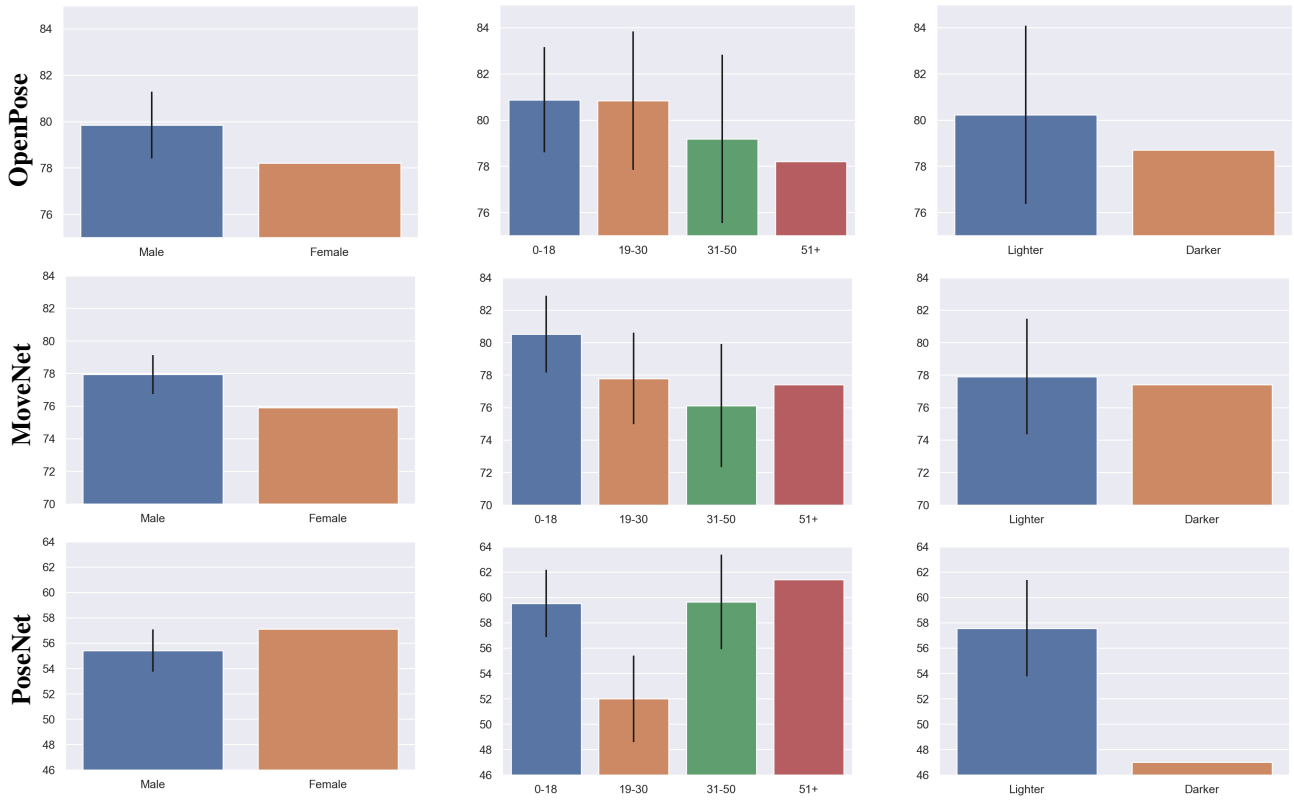
Takeaways. The analysis presented in this section aligns with the initial model card for MoveNet (Beletti et al. 2022), with the exception of the on-par performance on skin tone observed in Table 3c. This raises the question of why such a phenomenon is occurring, and more importantly, whether changing the split used for fairness evaluation could lead to different conclusion. For example, a model could be inherently discriminating against the female group, but the fairness evaluation could yield a very high performance because the set of images for the female group only contains easy-to-predict poses while the set of images for the male group contains difficult-to-predict poses. Towards the operationalization of fairness evaluations, we need to ensure the validity of our conclusions with respect to the available images.

6 Influencing Fairness Results

In this section, we explore whether the set of images used for fairness evaluation could influence the results. For example, one might select difficult examples for a specific sub-group to dampen its performance with respect to the other sub-groups. In turn, this would change the bias direction in the model and lead to invalid or ambiguous evaluations. It then becomes important to assess to what extent selected examples could influence the fairness evaluation.

Method

To evaluate the influence of the set of images in fairness evaluations, we subsample the over-represented groups in



(a) Gender (b) Age (c) Skin tone

Figure 5: **Adjusting sample size for sub-groups** on COCO-KD. We subsample sub-groups with a higher number of images (resulting in 223 images for both gender values, 56 images for every age group and 57 images for both skin tone values). We report the average and standard deviation of the mAP score over 100 different subsamplings. It is possible to find a subsampling set of images where the bias direction would change for every model.

COCO-KD to be of the same size of the group with the smallest size. This allows to provide a comparison with a similar sample size between groups. Concretely, we evaluate the same human pose estimation models as in Section 5: OpenPose (Cao et al. 2019), MoveNet Thunder (Chen et al. 2022) and PoseNet (Papandreou et al. 2018). Similarly, we consider the same demographic labels from COCO-KD: gender, age and skin tone; and we set them to have a sample size of 223, 56, and 57 images for each group respectively. This means that, for example, we have 223 images of both female and male genders and use the same set of images for all evaluated models. We perform the experiment 100 times, and report the average and standard deviation of the mAP score.

Results

Figure 5 presents the results of 2D human pose estimation when adjusting the number of images to be the same for every subgroup. When considering gender (*first* column), OpenPose and MoveNet still show a higher performance for males than females while there exist sets of images where PoseNet exhibits an on-par performance for both groups. When considering age (*second* column), trends are similar to Table 3b although results show large standard deviations. This means that it is hard to establish a performance ranking among the four different age groups. When considering skin tone (*third* column), it appears that the bias direction in OpenPose and MoveNet is very sensitive to the sample set, as the direction can show a discrimination against darker- or lighter-skins with a large margin by sampling the right subset of images. These results question the relevance of fairness evaluations, especially when dealing with an imbalanced dataset.

Takeaways. When a dataset for fairness evaluation presents an imbalanced distribution, it is interesting to evaluate a balanced scenario by subsampling the over-represented groups. If results end up contradicting initial evaluations, this means that fairness conclusions can be influences based on factors other than the demographic labels. For example, if a subgroup has easy-to-predict samples, then conclusions would not be meaningful. Other labels than the demographics attribute should then be collected to inform about how diverse and easy-to-predict the samples are. Towards the operationalization of fairness evaluations, we need to also assess whether the sample images for every subgroup are appropriate to measure the presence of potential biases.

7 Recommendations and Conclusions

In this paper, we have addressed the limitations and challenges related to operationalizing fairness evaluations in the context of a human-centric vision task. Through the case study of 2D human pose estimation, we have pinpointed the lack of available demographic annotations in all datasets. After semi-manually annotating a subset of the widely adopted COCO dataset for gender, age and skin tone labels, we observe how imbalanced the dataset is. We further evaluate commonly used 2D human pose estimation models and identify potential model biases. Given the low sample size for certain subgroups, we explore whether some sets of images could influence the fairness results. Building on this case study on 2D human pose estimation, we provide the following recommendations towards an improved operationalization of fairness evaluations:

Data collection. After reviewing the literature on human pose estimation, we noticed the lack of demographic annotations. When available, these are mainly done in a restricted context (*e.g.*, limited demographic labels or limited availability). This is an issue as such demographic annotations are mandatory to perform a fairness evaluation and understand potential data or model biases. As a result, such absence could harm the operationalization of fairness evaluations as practitioners may simple decide not to perform them or utilize inappropriate datasets. We recommend for future datasets for human pose estimation, and human-centric tasks in general, to collect demographic annotations from the start to facilitate fairness evaluations.

Annotations. While providing demographic annotations is crucial for fairness evaluations, it is also important to consider how these are collected. There exist multiple ways: automatic annotations from machine learning models; annotations by experts or crowdworkers; and self-reported annotations. We have observed in this paper that automatic annotations can be unreliable, as a manual check shows that a large proportion of the demographic labels can end up being incorrect. Having incorrect annotations could lead to invalid conclusions after a fairness evaluation. We recommend for future datasets to avoid automatic annotations, and rely on self-reported attributes or manual annotations or with manual quality checks to ensure the validity of the demographic labels.

Data imbalance. Access to demographic labels enables to explore how representative a dataset is. When annotating a subset of COCO with demographic labels, we noticed how imbalanced the dataset is, with male or light-skinned individuals being overly represented. This is a concern as such imbalance limits the operationalization of fairness evaluation on under-represented groups such as females, old-aged or darker-skinned individuals. This is particularly true for intersectional sub-groups, which can represent only a small fraction of the dataset, making it difficult to draw reliable conclusions. We recommend future dataset collection efforts to incorporate from the start a balanced demographic representation to enable more relevant fairness evaluations.

Fairness evaluations. Access to demographic labels also enables the exploration of biases in models. When relying on an imbalanced dataset for fairness evaluations, with potentially subgroups with a low number of samples, this creates additional challenges. Indeed, even if model bias is observed, it is unclear whether the phenomenon is meaningful or not. In our evaluations, we sub-sampled the over-represented subgroups to provide a clearer view of the presence of biases. For example, when considering MoveNet, we observe that the bias direction can easily flip either towards lighter-skinned or darker-skinned individuals. We recommend for future evaluations that rely on imbalanced demographics to report on multiple data splits to confirm the presence of model biases.

References

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *CVPR*.

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *CVPR*.

Andrus, M.; Spitzer, E.; Brown, J.; and Xiang, A. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *FAccT*.

AWS Rekognition. 2022. Detecting and analyzing faces. https://docs.aws.amazon.com/rekognition/latest/dg/faces.html. Accessed: 2022-08.

Beletti, F.; Chen, Y.-H.; Oerlemans, A.; and Votel, R. 2022. MoveNet Single Pose: Model Card. https://storage.googleapis.com/movenet/MoveNet.SinglePose\%20Model\%20Card.pdf. Accessed: 2022-08.

Bellamy, R. K.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilović, A.; et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5): 4–1.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *TPAMI*.

Cao, Z.; Martinez, G. H.; Simon, T.; and Wei, S. 2019. YA, Sheikh. Openpose: Realtime multi-person 2d pose, estimation using part affinity fields. *IEEE Transactions on Pattern, Analysis and Machine Intelligence*, 4.

Chardon, A.; Cretois, I.; and Hourseau, C. 1991. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science*, 13.

Charles, J.; Pfister, T.; Magee, D.; Hogg, D.; and Zisserman, A. 2016. Personalizing human video pose estimation. In *CVPR*.

Chen, Y.-H.; Oerlemans, A.; Belletti, F.; Bunner, A.; and Sundaram, V. 2022. MoveNet Thunder. https://tfhub.dev/google/lite-model/movenet/singlepose/thunder/tflite/float16/4. Accessed: 2022-08.

COCO Tasks. 2020. COCO 2020 Keypoint Detection Task. https://cocodataset.org/#keypoints-2020. Accessed: 2022-08.

Dantone, M.; Gall, J.; Leistner, C.; and Van Gool, L. 2014. Body parts dependent joint regressors for human pose estimation in still images. *TPAMI*, 36(11): 2131–2143.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Fabbrizzi, S.; Papadopoulos, S.; Ntoutsi, E.; and Kompatsiaris, I. 2022. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223: 103552.

Fabris, A.; Messina, S.; Silvello, G.; and Susto, G. A. 2022. Algorithmic Fairness Datasets: the Story so Far. *arXiv preprint arXiv:2202.01711*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *CVPR*.

Hanna, A.; Denton, E.; Smart, A.; and Smith-Loud, J. 2020. Towards a critical race methodology in algorithmic fairness. In *FAccT*.

Haque, A.; Peng, B.; Luo, Z.; Alahi, A.; Yeung, S.; and Fei-Fei, L. 2016. Towards viewpoint invariant 3d human pose estimation. In *ECCV*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

Holland, S.; Hosny, A.; Newman, S.; Joseph, J.; and Chmielinski, K. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy*, 12: 1.

Hossain, M. Z.; Sohel, F.; Shiratuddin, M. F.; and Laga, H. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6): 1–36.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *TPAMI*, 36(7): 1325–1339.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *ICCV*.

Johnson, S.; and Everingham, M. 2010. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *BMVC*.

Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; and Sheikh, Y. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *ICCV*.

Karkkainen, K.; and 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)*, 1548–1558.

Khalil, A.; Ahmed, S. G.; Khattak, A. M.; and Al-Qirim, N. 2020. Investigating Bias in Facial Analysis Systems: A Systematic Review. *IEEE Access*, 8: 130751–130761.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Landers, R. N.; and Behrend, T. S. 2022. Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*.

Lassner, C.; Romero, J.; Kiefel, M.; Bogo, F.; Black, M. J.; and Gehler, P. V. 2017. Unite the People: Closing the Loop Between 3D and 2D Human Representations. In *CVPR*.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2018. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. *arXiv:1812.00324*.

Li, J.; Wang, C.; Zhu, H.; Mao, Y.; Fang, H.-S.; and Lu, C. 2019. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*.

Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; and Li, Z. 2021. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In *CVPR*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Lin, T.; Dollár, P.; Girshick, R. B.; He, K.; Hariharan, B.; and Belongie, S. J. 2016. Feature Pyramid Networks for Object Detection. *CoRR*, abs/1612.03144.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; and Pietikäinen, M. 2020. Deep learning for generic object detection: A survey. *IJCV*, 128(2): 261–318.

Liu, S.; Yin, Y.; and Ostadabbas, S. 2017. In-Bed Pose Estimation: Deep Learning with Shallow Dataset. *arXiv:1711.01005*.

Liu, W.; and Mei, T. 2022. Recent Advances of Monocular 2D and 3D Human Pose Estimation: A Deep Learning Perspective. *ACM Computing Surveys (CSUR)*.

Madaio, M.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; and Wallach, H. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–26.

Madaio, M. A.; Stark, L.; Wortman Vaughan, J.; and Wallach, H. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Mehrotra, A.; and Celis, L. E. 2021. Mitigating bias in set selection with noisy protected attributes. In *FAccT*.

Minaee, S.; Boykov, Y. Y.; Porikli, F.; Plaza, A. J.; Kehtarnavaz, N.; and Terzopoulos, D. 2021. Image segmentation using deep learning: A survey. *TPAMI*.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *FAccT*.

Munea, T. L.; Jembre, Y. Z.; Weldegebriel, H. T.; Chen, L.; Huang, C.; and Yang, C. 2020. The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation. *IEEE Access*, 8: 133330–133348.

Nagpal, S.; Singh, M.; Singh, R.; and Vatsa, M. 2019. Deep learning for face recognition: Pride or prejudiced? *arXiv preprint arXiv:1904.01219*.

Papandreou, G.; Zhu, T.; Chen, L.-C.; Gidaris, S.; Tompson, J.; and Murphy, K. 2018. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*.

Parasurama, P.; and Sedoc, J. 2021. Gendered Language in Resumes and its Implications for Algorithmic Bias in Hiring. *arXiv preprint arXiv:2112.08910*.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 469–481.

Raji, I. D.; and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AIES*.

Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 33–44.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *IJCV*, 115(3): 211–252.

Sapp, B.; and Taskar, B. 2013. MODEC: Multimodal Decomposable Models for Human Pose Estimation. In *CVPR*.

Sigal, L.; Balan, A. O.; and Black, M. J. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1): 4–27.

Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; and Tian, Q. 2017. Pose-driven deep convolutional model for person re-identification. In *ICCV*.

Tang, R.; Du, M.; Li, Y.; Liu, Z.; Zou, N.; and Hu, X. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, 633–645.

Thong, W.; and Snoek, C. G. 2021. Feature and label embedding spaces matter in addressing image classifier bias. In *BMVC*.

Tompson, J.; Jain, A.; Lecun, Y.; and Bregler, C. 2014. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *NeurIPS*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, A.; Narayanan, A.; and Russakovsky, O. 2020. REVISE: A tool for measuring and mitigating bias in visual datasets. In *ECCV*.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*.

Wilson, B.; Hoffman, J.; and Morgenstern, J. 2019. Predictive inequity in object detection. *arXiv:1902.11097*.

Xiang, A. 2022. Being 'Seen' vs. 'Mis-Seen': Tensions between Privacy and Fairness in Computer Vision. *Harvard Journal of Law & Technology, Forthcoming*.

Xu, J.; Tasaka, K.; and Yamaguchi, M. 2021. Fast and accurate whole-body pose estimation in the wild and its applications. *ITE Transactions on Media Technology and Applications*, 9(1): 63–70.

Yao, B.; and Fei-Fei, L. 2010. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*.

Zhang, H.-B.; Zhang, Y.-X.; Zhong, B.; Lei, Q.; Yang, L.; Du, J.-X.; and Chen, D.-S. 2019a. A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5): 1005.

Zhang, S.-H.; Li, R.; Dong, X.; Rosin, P.; Cai, Z.; Han, X.; Yang, D.; Huang, H.; and Hu, S.-M. 2019b. Pose2seg: Detection free human instance segmentation. In *CVPR*.

Zhang, W.; Zhu, M.; and Derpanis, K. G. 2013. From actemes to action: A strongly-supervised representation for detailed action understanding. In *ICCV*.

Zhao, D.; Wang, A.; and Russakovsky, O. 2021. Understanding and evaluating racial biases in image captioning. In *ICCV*.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as Points. *CoRR*, abs/1904.07850.