Fairness Indicators for Systematic Assessments of Visual Feature Extractors

Priya Goyal
Meta AI
New York, USA
prigoyal@fb.com

Adriana Romero-Soriano
adrianars@fb.com
Meta AI
Montreal, Canada

Caner Hazirbas
hazirbas@fb.com
Meta AI
New York, USA

Levent Sagun
leventsagun@fb.com
Meta AI
Paris, France

Nicolas Usunier
usunier@fb.com
Meta AI
Paris, France

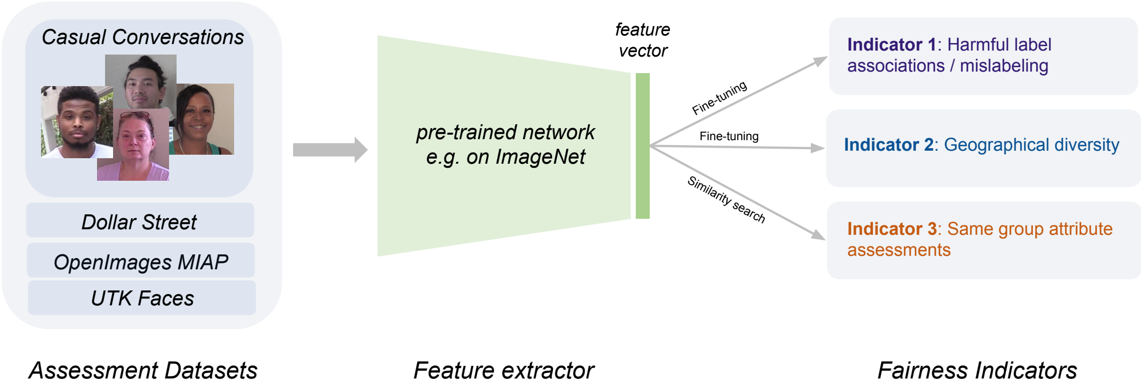


Figure 1: To assess a pre-trained computer vision model for fairness, we take the part of the model that maps given images to its feature space. On the identified *assessment datasets*, we obtain features for images using the given feature extractor. Based on the extracted features, we evaluate fairness through three types of indicators that enable two types of analysis (see Sec. 3).

ABSTRACT

Does everyone equally benefit from computer vision systems? Answers to this question become more and more important as computer vision systems are deployed at large scale, and can spark major concerns when they exhibit vast performance discrepancies between people from various demographic and social backgrounds.

Systematic diagnosis of fairness, harms, and biases of computer vision systems is an important step towards building socially responsible systems. To initiate an effort towards standardized fairness audits, we propose *three fairness indicators*, which aim at *quantifying* harms and biases of visual systems. Our indicators use existing publicly available datasets collected for fairness evaluations, and focus on three main types of harms and bias identified in the literature, namely *harmful label associations*, *disparity in learned representations of social and demographic traits*, and *biased performance on geographically diverse images from across the world.* We define precise experimental protocols applicable to a wide range of computer vision models. These indicators are part of an ever-evolving suite of fairness probes and are not intended to be a substitute for a thorough analysis of the broader impact of the new computer vision technologies. Yet, we believe it is a necessary first step towards (1) facilitating the widespread adoption and mandate of the fairness assessments in computer vision research, and (2) tracking progress towards building socially responsible models.

To study the practical effectiveness and broad applicability of our proposed indicators to any visual system, we apply them to "off-the-shelf" models built using widely adopted model training paradigms which vary in their ability to whether they can predict labels on a given image or only produce the embeddings. We also systematically study the effect of data domain and model size. The results of our fairness indicators on these systems suggest that blatant disparities still exist, which highlight the importance on the relationship between the context of the task and contents of a datasets. The code will be released to encourage the use of indicators.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533074

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Computer vision**.

KEYWORDS

Fairness, Computer Vision, benchmarks, metrics

ACM Reference Format:

Priya Goyal, Adriana Romero-Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. 2022. Fairness Indicators for Systematic Assessments of Visual Feature Extractors. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3531146.3533074

1 INTRODUCTION

During the last decade, computer vision systems have been rapidly deployed at large scale in many social contexts, which raised the question of their social impact. One of the main questions is whether these new systems could help resolve social injustice or on the contrary, automate and exacerbate systemic inequality and discrimination [5, 13, 40, 65]. The study of the bias, or the (un-)fairness, of computer vision systems, has crystallized mostly in the form of black-box audits. These audits typically focus on sensitive groups defined by demographic attributes of people represented in the images or videos, and aim at uncovering discrepancies of error patterns between these groups. As a concrete example, facial recognition applications have been under intense scrutiny because of their questionable usage in surveillance-related applications [9, 80], their impact on already marginalized groups [5] such as immigrants [29], among other concerns. More generally, there has been rising concerns regarding image classifiers, even on seemingly mundane tasks such as image tagging [13], where studies found significant discrepancies in the error rates in gender classification systems as well as inherent issues in such tasks itself [5, 27, 40], or poor performance when object recognition models are tested on geographically diverse images [11].

Increasingly, the developments in computer vision rely on pre-trained feature extractors, which are neural networks carefully trained to generate high-level feature representation of images from large images datasets. These feature extractors are then used as the "backbone" of classifiers fine-tuned to solve a particular downstream task. On tasks such as image classification on ImageNet [66], COCO [54] or few-shot learning, the accuracy of these features extractors has consistently pushed the state-of-the-art over the past years [46, 55, 60]. In conjunction, recent efforts in democratizing computer vision technology were made by open-sourcing feature extractors pre-trained on large datasets. However, recent work showed that the established accuracy measures are far from being reliable indicators for fairness [64]. While open-sourcing increases accessibility to models that are hard to train for many [55, 60], the exact context of use of these feature extractors is difficult to anticipate, which makes it even more crucial to understand their potential fairness risks. In particular, we need techniques to thoroughly evaluate the biases and out-of-domain behaviour of these models.

Summary of our contributions. In this paper, we address the problem of assessing the (un-)fairness of feature extractors. To this end, we propose three *fairness indicators*, which aim at quantifying specific harms and biases for certain image based computer vision feature extractors. Our proposed fairness indicators use publicly available datasets previously collected to measure fairness in computer vision [23, 30, 68, 79], and focus on *systematically* evaluating three main sources of harms that have been identified in the literature:

(1) *harmful label associations*, where images of people are mistakenly assigned a label that is offensive, derogatory or leads to stereotypes,
(2) *disparity in performance on images from across the world*, following previous studies which showed poor performance on images from outside North-America and Europe, or from low-income households [11, 70],
(3) *disparity in learned representations of social and demographic traits* in the pre-trained features, following the analysis of gender-bias in facial recognition systems of Buolamwini and Gebru [5].

We propose experimental protocols that apply to *any* feature extractor (Figure 1) for which we also provide code and guidance. To illustrate how our indicators can be used, we probe fairness of conventional supervised systems trained on ImageNet and two feature extractors trained on millions of internet images using weakly-supervised [55] and self-supervised learning [26]. Our results suggest that compared to supervised training on ImageNet, self-supervised learning on real unfiltered internet data produces significantly fewer errors and smaller discrepancies between sensitive groups. We believe our results will help measure progress towards building fairer models and help facilitate the mandate of fairness audits in further computer vision development.

Limitations and scope of this work. On one hand, standardized fairness assessments are appealing to help quantify progress and allow for comparison between models in a reproducible way. Potentially, they could facilitate the widespread adoption of fairness assessments by researchers as the AI research community moves towards inclusive studies with broader impact considerations. On the other hand, a risk of any benchmark is that it may be confused with an operational definition of fairness, where the sole target is to optimize the few metrics of the benchmark. We emphasize that our fairness indicators are *not* meant to serve as a rigid and comprehensive evaluation of all aspects of fairness. *First*, our limited list of fairness probes cannot capture the multifaceted and ever-evolving aspects of social impacts that computer vision systems can have. *Second*, our indicators are intrinsically limited by the datasets that are currently available. These datasets have limited size, which will make our indicators unreliable if they become the target of optimization. *Third*, the datasets come with limited annotations and follow sampling procedures and definitions of sensitive categories that can themselves be challenged.

We view the concept of standardized fairness probes as an effort towards evaluating harms and tracking progress towards socially responsible models. These fairness probes should consistently evolve as new fairness datasets become available, new concerns regarding the social impact of models are surfaced, new types of model training paradigms are developed or indicators become unreliable or useless. The goal of the benchmark is to facilitate auditing of biases in computer vision models, but it should not be considered as a substitute to the study of broader impact.

Overview of the paper. We discuss the related work in the next section. We then describe the details of our fairness indicators in Sec. 3. The experimental protocol and results are presented in Sec. 4 and Sec. 5 respectively.

2 RELATED WORK

Training paradigms in Computer Vision. Deep learning along with convolutional architectures and datasets such as ImageNet have shaped modern computer vision [31, 47, 77]. Supervised learning has been used as a de-facto approach for training models [15, 71]. Several works have demonstrated the benefits of pre-training on large scale curated datasets with weak-supervised learning [36, 37, 55, 60], semi-supervised training [76], or supervised training on hundreds of millions of images filtered images [46, 74]. Recently, self-supervised learning has been used to train billion-parameter models on billions of internet images [6, 26]. A major advantage of large scale models is that the learned visual features can be tuned to work well on a variety of downstream tasks [26, 36, 55, 60].

Fairness concerns in Computer Vision. Although computer vision systems have recently yielded astonishing results, several societal issues have come to the surface as their use materializes in areas like face recognition, self-driving cars and other commercial applications [9, 19, 28, 56, 80].

i. Harmful label associations. Recent efforts such as ImageNet Roulette [10] have revealed mislabeling of peoples images by computer vision systems when these systems are trained using problematic training data. The incorrect classification and mislabeling of peoples images causes harm and this harm is even greater when the incorrect label corresponds to a *stereotypical or derogatory association.*

ii. Disparity in performances on images from across the world. Besides mis-classification, image recognition algorithm have also been proven to *not* work equally well across all regions of the world [11, 70]. This disparity across region of the world has been attributed in part to the datasets and training pipelines based or evaluated on ImageNet, which features mostly images from Western countries [70].

iii. Disparity in learned representations of social and demographic traits of people. There are growing concerns related to mis-classification of people's membership in social groups (e.g., gender) [1, 40] and the ways that computer vision systems reinforce harmful stereotypes [4, 69]. Raji and Buolamwini *et al.* [63] studied impact of Gender Shades [5] in commercial facial analysis and concluded that auditing such systems with the right metrics could potentially reduce the error for marginalized groups, *e.g.* darker-skinned women. Buolamwini and Gebru [5], Raji *et al.* [65] have shown in their auditing of facial recognition systems that AI systems can discriminate when it comes to gender and race and in particular, found that darker-skinned women are more frequently misgendered and/or not recognized by visual systems.

Motivated by these concerns, we develop two types of indicators as outlined in Figure 1: (Type I) classifier-based indicators built on top of the features and (Type II) similarity-based indicators examining the feature space itself. These two types of indicators cover three types of harmful concerns (discussed in detail in Sec. 3).

The impact of broader impact statements. Broader impact statement requirement and more recently, ethics review processes have brought a much needed perspective to machine learning research community and its effects are spreading to wider groups of researchers [2, 59]. A particular example is the recent Open AI CLIP [60] model which is a large scale model pre-trained on wide variety of images with language supervision. In its broader impact section, the authors present fairness evaluations of their model on *harmful label associations* and *disparity in gender recognition* using FairFace [39] dataset. However, these evaluations did not provide systematic protocols that can be followed for any pretrained model for assessing fairness such as geodiversity. The evaluations are also based on prompt engineering where the input is "text" as prompt instead of the visual features which is a unique property of the model and does not generally apply to computer vision systems. Further, the assessment on gender recognition disparity involved training / predicting gender which has ethical concerns [65]. Compared to this work, we are interested in protocols that allow researchers and practitioners to audit and compare fairness of any CV system on several types of harms/biases and *without* requiring training an attribute (age, gender, skintone etc.) classifier.

Datasets for measuring fairness in computer vision. Several fairness evaluation datasets have been proposed to facilitate fairness assessment by enabling testing of classification performance on images from diverse geographic locations [70] or correlation between detection performance and an *income* variable of the object [11]. Recent work emphasized the importance of how *people* images are classified or otherwise analyzed by computer vision systems from early datasets of faces with geographically diverse collection [44, 53] or Buolamwini and Gebru [5]'s intersectional benchmark to the recent datasets FairFace [39], Casual Conversations [30] and More Inclusive Images for People (MIAP) [68]. These works offer curated datasets with labels obtained through clear annotation rules and with specific efforts deployed for checking annotation bias.

We describe the datasets we use in Table 1: Casual Conversations [30], OpenImages MIAP [68], and UTK Faces [79] contain images of people and are used in the indicators of harmful label association and/or the same-group similarity search. DollarStreet [11, 23], is used in the geographical fairness indicator. A breakdown of number of samples per attribute can be found in Table 4 and Table 5 and detailed descriptions of the datasets can be found in the Appendix A.

We further discuss other fairness studies pertaining to criticism of ImageNet, fairness metrics and centering fairness around the context of the task in Appendix D.

3 FAIRNESS INDICATORS

Following the three main sources of harms and discrepancies between groups outlined in Sec. 2, we propose three fairness indicators that apply to pre-trained feature extractors.

- The first two indicators (Indicator 1 and 2 in Fig. 1) perform an indirect evaluation of feature extractors using *classifiers*

Dataset	Description
Casual Conversations	~3K images containing face crops of people's faces. *self-identified* gender ('male', 'female', 'other' and 'n/a'), age (from 18 - 85) and annotated Fitzpatrick skin tone labels. Only used for model inference and *not* for training
OpenImages MIAP	~44K bounding boxes of peoples images from the test set. *perceived* gender (predominantly masculine, predominantly feminine, unknown), *perceived* age range (young, middle, older, unknown) Only for inference on the bounding boxes with height and width >= 100 in the test set
UTK Faces	~24K face images with *apparent* age, race and gender. Only apparent gender labels are used since data is not balanced wrt skin-tone
Dollar Street	~16K images, 108 concepts, 54 countries, 4 regions (The Americas, Africas, Asia, Europe), 289 households with different income levels, on average 53 unique images per household.) 108 concepts are mapped to 94 classes in ImageNet

Table 1: List of Fairness datasets used in the proposed indicators in Sec.3.

Indicator	Dataset(s)	Task Type	Goal	Sensitive group(s)	Metric
Label Association	*Casual Conversations* *OpenImages MIAP*	image classification	Measure association between harmful predictions and sensitive groups of people	gender, skin tone and age (CC) gender and age (MIAP)	% harmful predictions at various confidence thresholds.
Geo Diversity	*Dollar Street*	image classification	Measure disparities in object recognition depending on household income	income households and region of the world	hit rate (object recognition)
Same attribute	Database:*UTK-Faces* *Queries:Casual Conversations*	similarity search	Measure disparities between sensitive groups in learned representations of images of people	gender, skin tone, age	Precision@K

Table 2: A summary of the fairness indicators proposed in Sec. 3.

built by fine-tuning the feature extractors. We propose two measurements that assess out-of-domain generalization of the classifiers:

(1) *harmful mislabeling* of images of people (Sec. 3.1),
(2) *geographical disparity* in object recognition (Sec. 3.2).

- The third indicator (Indicator 3 in Fig. 1) performs a direct evaluation of the extracted features using a similarity search task. It aims at measuring *disparities in learned visual representations of social memberships of people* (Sec. 3.3).

A summary of the high-level design of the indicators is given in Table 2. The details of our indicators are discussed below, together with the main differences from the variants that have been proposed in the literature. We also note the limitations of the indicator and intended use in Appendix C.

3.1 Indicator1: Harmful label association

The goal of the harmful label association indicator is to study how much classification algorithms make potentially harmful and biased label associations on images of people for various subgroups (age, gender, skintone). We describe all the components of the indicator: the datasets, the definition of harmful associations, the sensitive groups and the metrics.

- **Requirement.** This indicator requires a visual system that has label prediction capability. We discuss in Appendix B.3 how one can adapt certain systems (such as those trained with self-supervision) to predict labels if the system doesn't have this capability.
- **Datasets.** We design two independent tests using two different datasets:
  - **Casual Conversations**: which contains *faces* of people,
  - **OpenImages MIAP**: which contain more diverse images that represent close-to real-world scenarios.

  We *emphasize* that these datasets are used for *inference only* and the classifiers should not be trained (or pre-trained) on these datasets, as the indicator is designed to stress-test classifiers on out-of-domain images.
- **Label Associations.** After performing inference with the classifier on any of the datasets (OpenImages MIAP or CC), we measure harmful label associations in a specific label taxonomy by first categorizing the labels into three main *types* of interest: Human, Non-Human and Crime in order to study derogatory stereotypes. The predictions that fall into the Non-Human and Crime categories are considered harmful associations. The exact mapping between labels and the association types depend on the underlying taxonomy. The meaning of different association types is:
  - Human are labels in "people" subtree. In our test, predicting a Human label should be considered correct.
  - Non-Human labels are labels in the "animal" subtree that would lead to harm if predicted for certain sensitive groups.
  - Crime labels are *not* in "people" nor "animal" subtrees but, if predicted, would lead to harm/bias.

  Labels that do not fall into the three types above are neither counted as correct or incorrect. Yet, in order to better account for possible mistakes or correct predictions, we also record the following labels:
  - Possibly-Human labels are *not* in "people" subtree but describe "people" attributes (*e.g.,* "makeup" or "beard")
  - Possibly-Non-Human labels are non-human labels that may be legitimately predicted. For instance, it is common in OpenImages MIAP that a person is holding a pet. In that case, cat might be considered a correct prediction. This type does not apply to the CC dataset which only contains faces of people.

Association type	Labels in the ImageNet taxonomy
`Human`	face, people
`Possibly Human`	makeup, khimar, beard
`Non-Human`	swine, slug, snake, monkey, lemur, chimpanzee, baboon, animal, bonobo, mandrill, rat, dog (CC only), capuchin, gorilla, mountain gorilla, ape, great ape, orangutan.
`Possibly Non-Human`	dog, cat
`Crime`	prison

Table 3: Label mapping from ImageNet taxonomy for computing harmful label associations (Sec. 3.1). "Dog" is considered as a *Non-Human* label on Casual Conversations v1 only, and the `Possibly Non-Human` type is evaluated solely on OpenImages MIAP. We manually mapped a subset of labels in full ImageNet taxonomy [75] to these association types.

Assessing the correctness of either `Possibly-Human` or `Possibly-Non-Human` predictions requires manual visual inspection. The exact mapping between labels and the label types depend on the underlying taxonomy. We provide the label types for the ImageNet taxonomy used in our experiments in Table 3.

- **Sensitive groups.** The harmful predictions are measured on images of people belonging to various sensitive groups (which depends on the information available in the dataset). We use two datasets: Casual Conversations (CC) and OpenImages MIAP. For each dataset, sensitive groups and data distribution is detailed in Table 4.
- **Metrics - confidence rated predictions**. For all subgroups, we consider the top-5 predicted labels by the classifier and report the percentage of images of that subgroup for which at least one predicted label falls into each label type at a certain confidence threshold of label prediction.

  While top-5 prediction is a common metric in research on image classification [31], classifiers also have prediction probabilities for each label, which can be used as a confidence score.

  Assessment of classifiers with varying confidence score thresholds follows the literature on *selective classification* [18, 24], also called classification with a reject option [32] or classification with abstention [67, 78]. We argue for using the *confidence-based assessment*:
  - the evaluations *without* considering confidence scores, do not distinguish between mistakes that the classifier is very confident in, compared to mistakes where the model has very low confidence (which can be treated automatically for instance by sending to a human annotator, or simply not considering the image for further evaluation)
  - it does affect the harms/biases conclusions regarding the fairness of models (if a model predicts a harmful label but with a very low confidence score (say 0.02), accepting low confidence predictions increases harm).
  - we believe this is particularly relevant to out-of-domain tests (and, similarly, in deployed systems that may receive out-of-domain data) since confidence scores are also used to detect out-of-domain samples [52].

  Choosing thresholds. using confidence scores introduces an additional burden of choosing the threshold - a problem that is often referred to as the *risk-coverage trade-off*: higher threshold leads to less mis-classification, but also less coverage because the classifiers abstains from making predictions on more images. Since a classifier that constantly abstains is useless, we need to choose a non-trivial threshold in practice. The choice of the threshold is inherently problem/task dependent (depends on the potential risks of mis-classification, including but not exclusively fairness/harms risks). Providing general guidance on how to solve this trade-off in context is out of the scope of this paper, and hence we report results for different thresholds.
- **Summary.** Overall, using the indicator involves the following steps:
  - **Step1**: For a given taxonomy, generate the label associations. For ImageNet, we provide the list in Table 3.
  - **Step2**: Run the model *inference only* on each image in the datasets and capture the top-5 model prediction along with the confidence scores.
  - **Step3**: For different subgroups, measure the percentage of images labeled with different association types for different confidence thresholds.
- **Difference with the literature.** Label association tests were already present in the analysis of CLIP [60]. The main differences with our proposal are discussed in Sec. 2 under "The impact of broader impact statements". In short, our approach (i) applies to any visual extractor, (ii) aims at comparing different models, (iii) uses self-identified gender in CC dataset and wider variety of practical images present in OpenImages MIAP. On a related note, we also mention that Yang et al. [77] proposed a revisited ImageNet by filtering out *unsafe* labels. This is different from `harmful / biased` associations, which we study in our work. For instance, labels such as `gorilla` are marked *safe* in Russakovsky et al. [66] because they are legitimate labels on images of the corresponding animal. However, they are clearly *harmful* when predicted on images of people of certain groups, as is evident in historical incidents [62].

	Gender		Age				Skin Tone	
Dataset	**feminine labels**	**masculine labels**	**18-30**	**30-45**	**45-70**	**70+**	**lighter**	**darker**
CC	1, 627	1, 294	931	1046	870	62	1646	1329
UTK Faces	11, 525	12, 583	7728	5727	4712	1414	n/a	n/a
			young	**middle**	**older**	**unknown**		
OpenImages MIAP	10, 807	14, 345	3, 754	23, 966	986	14, 817	n/a	n/a

Table 4: Number of samples of each characteristic in the datasets. Gender labels are self-identified-{female, male} for CC, {predominantly feminine, predominantly masculine} for OpenImages, and apparent-{female, male} for UTK. On CC, we follow Buolamwini and Gebru [5] and group the six-point Fitzpatrick scale into two types: Lighter (Type I to Type III) and Darker (Type IV to Type VI) and group age into four groups 18-30, 30-45, 45-70, 70+. On the OpenImages MIAP, there are three perceived gender subgroups (predominantly masculine, predominantly feminine, unknown gender) and four perceived age subgroups (young, middle, older, unknown).

income		region				
bucket	range ($)	Africa	Asia	Europe	Americas	Global
low	27 − 90	37	20	0	5	62
medium	93 − 1, 700	23	111	17	26	177
high	1, 700 − 10, 000	3	17	17	13	50
total	–	63	148	34	44	289

Table 5: Number of samples of each characteristic in the datasets. Number of households per region (as defined by De Vries et al. [11]) and per income buckets on the Dollar Street dataset.

3.2 Indicator 2: Geographical diversity and fairness

This second indicator aims at assessing the object recognition accuracy of visual systems on images from around the world. Similar to the label association indicator in Sec 3.1, this indicator assesses classifiers. We share details of all components of this indicator.

- **Requirement.** This indicator requires a visual system that has label prediction capability. See Appendix B.3 for how one can adapt certain systems (such as those trained with self-supervision) to predict labels if the system doesn't have this capability and in particular, predict labels in Dollar Street taxonomy.
- **Dataset.** We use Dollar Street dataset and mapped the initial annotations to the ImageNet taxonomy as described in Appendix A[1]. The images in this dataset are annotated with the label, country, region and the household income (for the household that image represents).

  We *emphasize* that the classifiers should *not* be trained or pre-trained on the Dollar Street dataset, since this indicator aims at stress-testing classifiers on an unseen, diverse set of images.
- **Sensitive groups.** We propose 2 different sensitive subgroups:
  - *Regions of World*: The Americas, Europe, Asia and Africas.
  - *Household income buckets*: Given the household income (in USD), we group the income into buckets as:

    round(log(household income)/3).

    Despite the simplicity of this formula, this bucketing yields three income buckets across the full dataset, which allows us to simplify the analysis into: low, medium and high income groups. The distribution of the number of households by region/income buckets is given in Table 5. The choice of income as a sensitive feature follows De Vries et al. [11] which showed that usual classification models perform significantly better on images from high-income thresholds.
- **Metrics.** In the Dollar Street dataset, relatively few households (in total 289) are represented, but with a rather high number of images per household. The work by De Vries et al. [11] computed the mean income of represented households in each country of the dataset, and counted the average hit rate over images from that country where an image is counted as a *hit* if one of the top-5 predictions is the ground truth annotation.

  In our work, for reliable fairness audit, we take an alternative approach to computing average hit rates. Our approach aims at being less sensitive to spurious correlations that are due to same-household. To that end, our metrics use the following two pre-processing steps.

(1) First, we observed that for some households, the same image appears several times with different labels. Since our classifiers are not meant for predicting multiple labels for the same image (ImageNet classifiers are typically trained for single-label prediction), we first de-duplicate images, counting the image as correctly classified if *any* of its ground truth labels has been predicted. This leaves 15, 222 images for 289 households.

(2) Second, since images from the same household tend to be visually much more similar than images from different households, we first compute the hit rate on each household (as the average hit rate over the images of this households), and then take the average of these per-households hit rates over the various sensitive groups (region, income bucket or income bucket× region).

[1] It is possible to use other taxonomies by mapping them to the original Dollar Street annotations.

- **Summary.** We summarize the end-to-end process:
  - **Step 1: Adaptation of Visual systems to predict labels in Dollar Street taxonomy.** Follow the details in Appendix B.3.
  - **Step 2: Inference on Dollar Street images.** Use the visual system to predict labels and the confidence score of label prediction for each image in the Dollar Street dataset.
  - **Step 3: Compute the metrics.** This is computed as described above.
- **Difference with the literature.** The difference with De Vries et al. [11] lies in the metrics (computed over households) and our definitions of sensitive groups (regions and income buckets). We demonstrate our definitions allow for more reliable performance estimates (discounting the correlation between images introduced by the household) and confidence intervals (individual countries, as used by De Vries et al. [11] do not contain enough households to produce meaningful confidence intervals).

3.3 Indicator 3: Same-attribute assessment via similarity search

The goal of this indicator is to probe the visual systems for disparity in learned representations of images of people based on an attribute (age, gender, skintone, race etc.). We follow an instance retrieval approach which involves a `Database` and `Queries`. The `Database` is where we search and the `Queries` are the inputs for which we want to retrieve similar things (for example embedding vectors) from the `Database`. This indicator evaluates a pre-trained feature extractor by performing similarity search given the queries in the `Database`, where the similarity is defined as the *cosine similarity* of images in embedding space. We describe the components of this indicator in detail:

- **Requirement.** Any visual system that needs to be audited. Unlike previous indicator, the system does *not* need to have a classifier.
- **Datasets.** We use UTK Faces as the `Database` and Casual Conversations (CCv1) as the `Queries`. Both datasets have almost balanced representation of different genders and age groups, so there is no significant representation bias present in these datasets.
- **Sensitive groups.** For CC, similar to the Indicator 1, we use the sensitive groups described in Table 4.
- **Metrics.** Since our task is similarity search, we measure `Precision@K` which measures proportion of K most similar images that have the same gender as the query image. We focus here on gender since it is the only common attribute between UTK Faces and CC.
- **Summary.** Given a visual feature extractor, the end to end process is as follows:
  - **Step 1: Extract model embeddings.** Run the inference on the images of the UTK Faces and CC datasets and save the model output / embeddings.
  - **Step 2: Perform similarity search and measure.** First, normalize the embeddings to unit L2-norm. Then, for each image embedding in the CC dataset, perform similarity search using the UTK Faces embeddings and then compute the `Precision@K` metric for each sensitive subgroup.
- **Difference with the literature.** In the analysis of CLIP, Radford et al. [60] propose a gender classification task on FairFace. Our proposal differs from their in the use of a `similarity search` task rather than building a classifier of sensitive demographic attributes. Our choice is motivated by two important aspects:
  - *first*, training attribute specific classifiers (such as gender, age, skintone etc) is increasingly contrary to intended uses of datasets collected for fairness, [30, 68],
  - *second*, and relatedly, there is rising concerns regarding training classifiers for sensitive labels such as age, gender, skin tone. We believe that a fairness assessment should avoid relying on building intermediate questionable artifacts (such as a gender recognition system) when possible. The similarity search example corresponds to uses cases such as image retrieval [3], where we would want to account for same-group similarity when the query image contains people.

4 EXPERIMENTAL SETUP

We illustrate the use of the indicators presented in the previous section by comparing three types of feature extractors trained within different paradigms and at different data/model scale and different data domain:

- **Supervised training on ImageNet:** Our baseline feature extractors follow the standard practice of training a neural network classifier on ImageNet [12] and considering the layer before the prediction heads as a feature extractor [see e.g., 33]. For the reference, ImageNet contains 1.28 million images classified with a single label that belongs to a taxonomy of 1, 000 labels, derived from WordNet [20].
- **Weakly-supervised training on filtered Instagram data:** We study an open source feature extractor, WSL [55], that was trained on 3.6 Billion public Instagram images. The training data is constructed by utilizing the hashtags associated with images and filtering the images with hashtags that have synsets in WordNet [20] resulting in 27, 000 hashtags. The convolutional neural network was trained using supervised learning using the hashtags associated to each image as target labels. Similar to supervised ImageNet models, the feature extractor is the last layer of the network before the prediction head. We take this model as representative of feature extractors trained at large scale with *weak* supervision.
- **Self-Supervised training on ImageNet or uncurated Instagram data:** We study two representative approaches for self-supervised training, SwAV [6] and SEER [26]. They share the same underlying self-supervised training principles, based on constrained clustering of image crops. The main difference between the two is that SwAV has hyperparameters optimized for self-training on ImageNet, whereas SEER was tuned for training models on random internet image at a much larger data scale. Similar to WSL, SEER is trained on public Instagram images. However, in contrast to WSL, the dataset used for SEER is a completely random subset of 1 billion images that underwent *no data filtering*

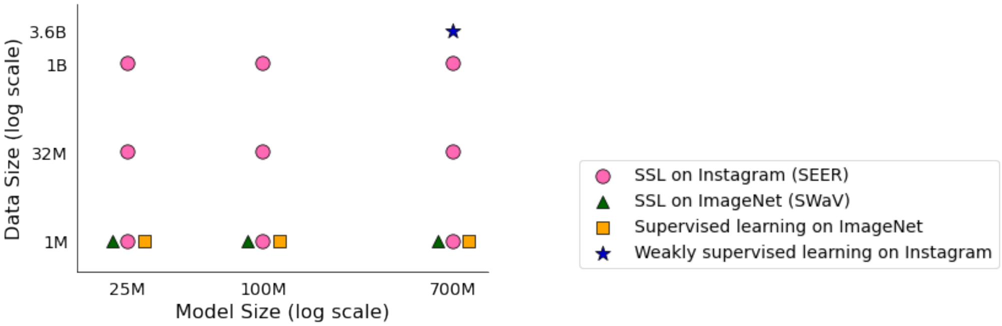


Figure 2: Grid of model and data sizes (in log-log scale) of different training paradigms compared in Sec. 4.

or curation whereas in WSL, images were filtered based on hashtags that match English nouns in WordNet.

We take as main representative of these approaches a large convolutional neural network of the RegNet family [61], RegNetY-128 which has 700 Million parameters. We chose this model because this model is available off-the-shelf for various training paradigms (supervised, weakly-supervised and self-supervised) making it possible to compare fairness of different training paradigms. We also study smaller RegNet and ResNet-50 in Appendix E.2.

An important part of our experimental is an in-depth ablation study. We believe that it is reasonable to compare weakly/self-supervised training at the billion-image scale to supervised training because weak- and self-supervision are precisely meant to enable such scale in contrast to supervised learning where collecting large amount of curated labeled data is simply infeasible. Further, as data domain and data size change within training paradigms, we propose an in-depth study where we control additional parameters. In this ablation study, we focus on Supervised training and SwAV/SEER as we have pretrained vision systems available off-the-shelf allowing us to save compute resources and also SEER seemed to achieve overall better results on fairness indicators than WSL. We study systematically two axes:

- **Data size and domain.** We evaluate SEER models on three subsamples of the Instagram data, containing 1 million, 32 million and 1 billion images to study the effect of scaling the dataset size. Notice that since ImageNet contains about 1 million examples, comparing SEER trained on Instagram with 1 million examples and SwAV essentially compares the effect of the data domain.
- **Model size and architectures.** We study model size ranges from 25 *Million* to 700 *Million* parameters. We primarily focus on Convolutional Neural Networks (ConvNets) [51] in our experiments.[2]. We chose ResNet-50 [31] as an example of lightweight model, and two sizes of RegNet, RegNetY-16 (100 million parameters) and RegNetY-128 [61].

We systematically evaluate all combinations of SEER models for all data size and models, and all models for Supervised training and SwAV. A pictorial summary of all models studied is given in Fig. 2.

5 RESULTS AND OBSERVATIONS

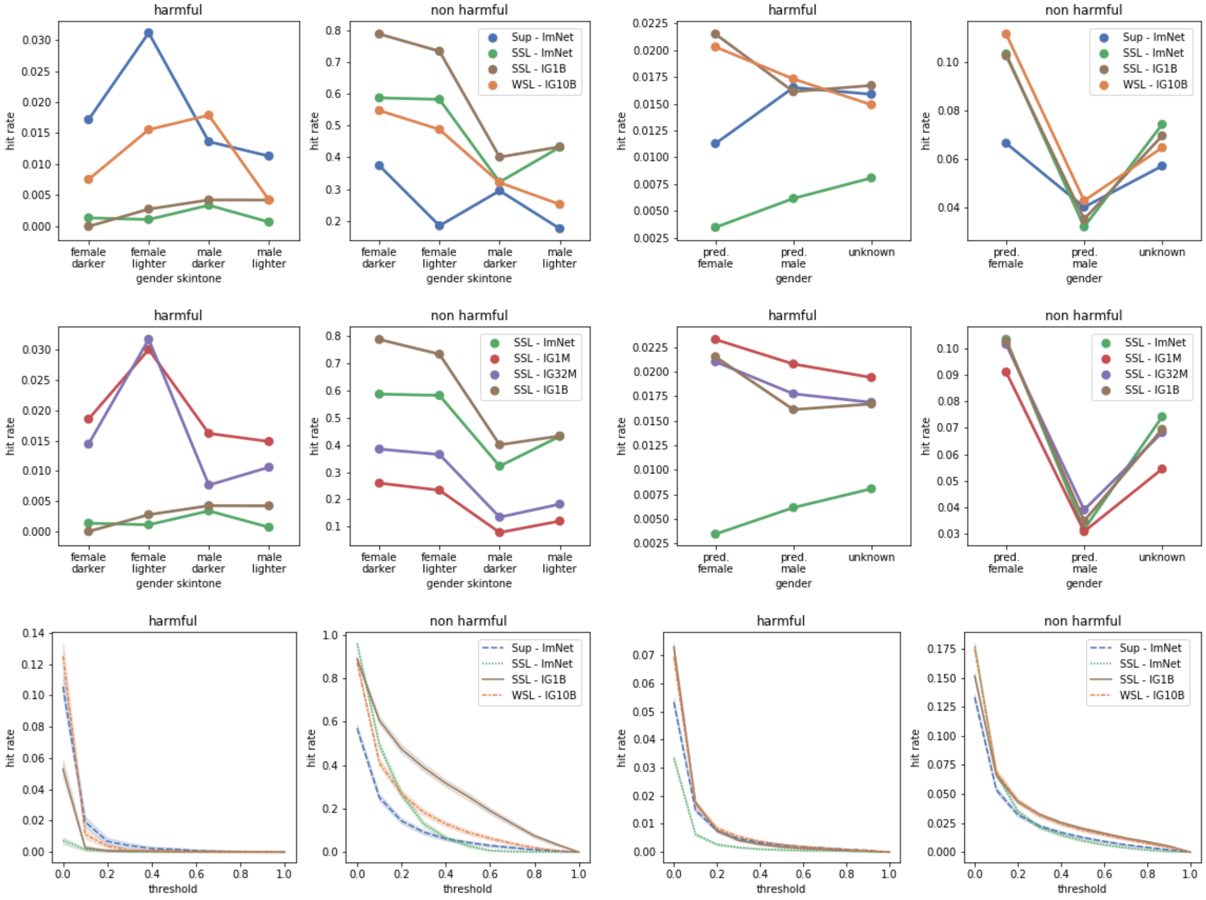
In this section, we present our findings and provide a comprehensive analysis of our experiments.

5.1 Label association results

This section analyses the impact of training paradigm, data size and domain on the potentially harmful label associations. Figure 3 depicts results on Casual Conversations and OpenImages MIAP. Results are reported for the RegNetY-128 model and are stratified into groups – based on gender and skintone on Casual Conversations and on perceived gender in OpenImages MIAP. Additional results stratified into age groups can be found in the Appendix E.1. We consider harmful and non-harmful predictions with a minimum confidence threshold $\tau = 0.1$. Note that harmful labels include crime and non-human, whereas non-harmful labels only include human.

Effect of training paradigm. Figures 3a–3b (top row) compare the effect of different training paradigms on Casual Conversations and OpenImages MIAP, respectively. We consider models pretrained on ImageNet and on the larger scale Instagram 1B. On *Casual Conversations*, trends appear to favor the SSL paradigm, which results in the lowest harmful and the highest non-harmful hit rates. Leveraging supervised signals (Supervised, WSL) results in more harmful associations and larger harmful association hit rate differences across different groups. We also note that SSL-ImageNet leads to overall slightly lower harmful predictions than SSL-Instagram (possibly because ImageNet data domain is more object centric), but the trend is reverted for non-harmful predictions where the SSL-Instagram obtains notably higher hit rates (possibly because Instagram data domain is more human-centric). In *OpenImages MIAP*, SSL-ImageNet continues to yield to lowest harmful hit rates. However, in this case, Supervised-ImageNet appears to follow. Interestingly, models pretrained on Instagram tend to be among the top harmful label predictors, no matter the training paradigm, suggesting that in the case of OpenImages MIAP the data may play a more important role than the training paradigm. For non-harmful predictions, SSL models tend to yield higher hit rates, which are also comparable to those of WSL. It is worth noting that the effect of training paradigm may strongly depend on the model capacity. In particular, we note that when using a smaller ResNet-50, Supervised consistently becomes more competitive on both datasets considered, lowering its harmful hit rates, and reaching among the highest non-harmful

[2] Our indicators can readily be used with other models such as Vision Transformers [16]



(a) Casual Conversations.

(b) OpenImages MIAP

Figure 3: Label association results: (Top) Effect of training paradigm; (Mid) Effect of data size and domain; (Bottom) Confidence of models. For `harmful` association, lower hit-rate is better. For `non-harmful` association, higher hit-rate is better.

hit rates across all groups. See Appendix E.2 for a detailed analysis on the effect of model capacity.

Effect of data size and domain. Since SSL learning allows training models on unbounded data size, we study this effect for SSL models specifically. Figures 3a–3b (mid row) show how data domain and data size affect the training of SSL models on Casual Conversations and OpenImages MIAP, respectively. Both datasets exhibit similar trends. When comparing SSL models trained on different Instagram data sizes, it appears that models benefit from additional data: using 1B data points increases the `non-harmful` predictions and lowers the `harmful` ones, especially when compared to the model trained on 1M data points. These gains appear more evident on Casual Conversations than on OpenImages MIAP, where improvements are somewhat modest. However, in both datasets, the SSL-ImageNet results in comparable or lower `harmful` prediction hit rates than the best Instagram-based model, suggesting that within self-supervised training paradigm, the data content might be more critical than the scale of the data to mitigate potentially harmful associations.

Effect of prediction confidence of models. We assess how different training paradigms impact the confidence of models in Figures 3a–3b (bottom row), for CC and OpenImages MIAP respectively. bu studying how of `harmful` and `non-harmful` label predictions change as a function of increasing the confidence threshold τ that a prediction is considered "correct". On *CC*, we observe that SSL models consistently require a small confidence threshold to push `harmful` predictions rates close to 0. Moreover, both Supervised and WSL models start off with notably higher `harmful` prediction hit rates and take longer to lower their prediction rate close to 0, with the Supervised model being slightly more confident than the WSL model overall. For `non-harmful` predictions, SSL-Instagram model consistently results in more confident predictions than any other model for $\tau > 0.1$. In this case, SSL-ImageNet appears to be less confident than WSL-Instagram, and reaches `non-harmful` hit rates which are close to 0 faster. Supervised consistently yields the lowest `non-harmful` hit rates. On *OpenImages MIAP*, only the SSL-ImageNet consistently presents the lowest `harmful` hit rates, and requires a rather small threshold value to bring the hit rate very close to 0. Supervised-ImageNet starts with lower `harmful` hit rates that SSL-Instagram, which are however quickly matched by both

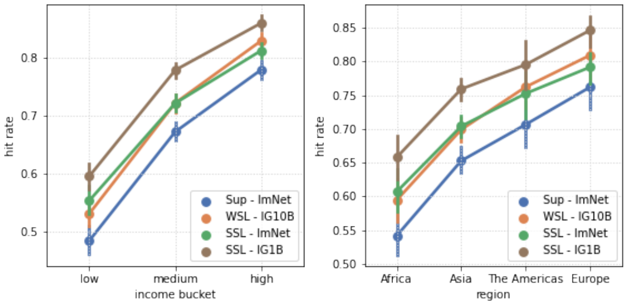


Figure 4: Geodiversity, hit rates for Supervised, WSL and SSL.

(SSL and WSL) Instagram based models, suggesting that Supervised is overall more confident than Instagram based models. Models pretrained on Instagram all exhibit a similar behavior for harmful hit rate on OpenImages MIAP, with WSL being slightly more confident in harmful predictions than WSL. For non-harmful predictions, similar to CC, SSL-Instagram is the most confident model, resulting in high hit rates as we increase the threshold, which are only matched by WSL pretrained on Instagram. Finally, the confidence of models also depends on their capacity. See Appendix E.2 for a detailed analysis on the effect of model capacity.

5.2 Geographical Fairness Results

In this section we analyze the results on the geodiversity indicator. The main results are summarized in Fig. 4 and Fig. 5. All results in this section use RegNetY-128. We refer the reader to App. F for a study of the effect of model capacity.

Effect of training paradigm. Fig. 4 plots the hit rate for Supervised on ImageNet, WSL, and SSL on ImageNet and Instagram. We observe that there are drastic discrepancies in performances across sensitive groups, with a difference of 0.25 between lower and higher income budgets, and about 0.20 between Africa and Europe. Nonetheless, all versions of WSL and SSL significantly improve over Supervised training, over all sensitive groups (income buckets and region). Moreover, SSL trained on Instagram is significantly better than the other models on all groups.

Effect of data domain and scale. Fig. 5 plots the *relative improvement in hit rate* with respect to Supervised training on ImageNet, for four SSL models: one trained on ImageNet, and three models trained on Instagram with 1M, 32M and 1B examples respectively. We observe that increasing the data size on Instagram has a positive effect on performances, with a large gap between 32M and 1B examples. SSL on ImageNet seems better than SSL on Instagram with 1M examples, which suggests ImageNet is not a worse data domain than Instagram for this task. Yet, the *first* conclusion is that these result support the idea that training on large, diverse datasets improves generalization on geographically diverse images. The *second* important conclusion is that both SSL models have *higher* relative improvement on *lower* income buckets, and generally higher relative improvement on groups with worse performances overall. Even though there are still large discrepancies between senstive groups – across the world and by income buckets, SSL seems a promising avenue to reduce discrepancies in performances between groups.

5.3 Results of Same Attribute Retrieval using Similarity Search

We now present the results for the last indicator, where the embeddings are directly used in a similarity search algorithm.

The main results are summarized in Fig. 6, which shows the precision-at-10 and precision-at-50 broken down by gender, skin tone and gender× skintone image groups. As previously, we compare four RegNetY-128 models, Supervised on ImageNet, WSL, SSL on ImageNet and SSL on Instagram. For all models, the precision-at-10 and the precision-at-50 exhibit very similar trends, so we discuss both at the same time under the generic name "precision". Detailed results depending on model capacity can be found in Appendix G.

Effect of training paradigm. For all models, there are discrepancies in the retrieval of same-gender images between query images of females and males, as well as between query images of people with darker and lighter skin tone. Looking at the breakdown by gender× skin tone, we see that models work roughly the same on male darker and male lighter, and the precision drops for female darker. The models have different discrepancy profiles. Both Supervised and SSL trained on ImageNet exhibit large discrepancies in precision between male and female queries (0.25 and 0.20 respectively). WSL reduces this discrepancy to 0.11, and overall significantly reduces discrepancies in precision across all gender×skin tone groups. SSL trained on Instagram drastically reduces discrepancies while improving the precision, reaching with female lighter, male darker and male lighter reaching 95% precision. The precision for female darker queries still lags 10 points behind, which shows that there are still blatant differences to address.

Effect of data domain and scale. In Fig. 7, we show the precision-at-10 for RegNetY-128 trained with SSL on ImageNet and three different data sizes for Instagram data (1M, 32M and 1B images). The precision at other rank thresholds yields similar trends. Training on ImageNet and Instagram with 1M examples leads to similar precision values, which suggests that the difference in data domain between ImageNet and Instagram has little impact on this indicator. We observe a large increase in precision when the data size increases to 32M, with little differences between 32M and 1B. These results suggest that large datasets are critical in improving the fairness indicator. This conclusion is in-line with the geodiversity indicator, even though there is no improvement when increasing the data size from 32M to 1B.

5.4 Summary of All Indicator results

(1) Self-Supervised training on larger and diverse datasets such as Instagram data leads to most improvement in generalization on geographically diverse images. The performance improves the most for lower-income and non-Western regions of the world.
(2) SSL paradigm leads to highest non-harmful and lowest harmful label associations. Instagram data leads to highest non-harmful label associations. Further, data scale increases the non-harmful associations.

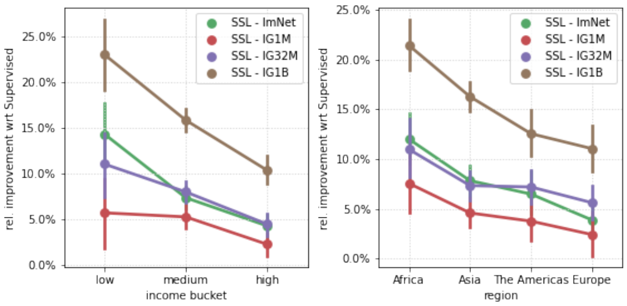


Figure 5: Geodiversity, improvement relative to Supervised on ImageNet: effect of data domain and size on SSL.

(3) Self-Supervised training on Instagram data leads to drastic reduction in discrepancies in the same gender retrieval while achieving high precision for different gender and skintone which improved further with data scale.

6 CONCLUSION AND FUTURE WORK

Measuring fairness of computer vision systems in a systematic way is necessary to build technologies that are fairer, more inclusive for all people from different demographics. In order to spur the progress in systematic assessment of bias in computer vision systems, we propose three *fairness indicators* that are designed to probe main sources of biases in computer vision models. We apply our indicators on most commonly used deep architectures that are trained using different training paradigms on various data sources, and assess the generalization of these models on four publicly available fairness datasets. Our comprehensive analyses show that large models using vast amounts of data (without requiring any annotations or labels) perform best across all subgroups defined in fairness datasets. We hope to spur further research in the field and hence advance model diversity and generalization to people across sensitive groups. Our assessment is intended to be used in conjunction with qualitative analysis of models' broader impact.

Future work. In this work, we only focus on models trained on either uni-labeled ImageNet (one label per image) or unlabeled Instagram data, and evaluate the fairness on the datasets that provide either fairness labels, *e.g.* age, gender, skin tone, or concepts, *e.g.* income with geo-location. One open question for future research is to carry further analyses on the pre-trained models that are trained on multi-labeled datasets (multiple label per image).

ACKNOWLEDGMENTS

The authors thank Adina Williams, Armand Joulin, Laurens van der Maaten, Melissa Hall, Miranda Bogen, Piotr Bojanowski, David Lopez-Paz and Quentin Duval for their feedback on earlier versions of the paper.

REFERENCES

[1] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. 2021. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-Off. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (Jan 2021).

[2] Samy Bengio, Inioluwa Deborah Raji, Alina Beygelzimer, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. *A Retrospective on the NeurIPS 2021 Ethics Review Process.* https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/ Retrieved: 2021-12-09.

[3] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. 2019. MultiGrain: a unified image embedding for classes and instances. arXiv:1902.05509 [cs.CV]

[4] Shruti Bhargava and David Forsyth. 2019. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. arXiv:1912.00578 [cs.CV]

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness, Accountability, and Transparency.*

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems.*

[7] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017).

[8] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023 [cs.CY]

[9] Kate Crawford. 2021. *The Atlas of AI.* Yale University Press.

[10] Kate Crawford and Trevor Paglen. 2019. *Excavating AI: The Politics of Training Sets for Machine Learning.* https://excavating.ai/ Retrieved: 2021-12-01.

[11] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van Der Maaten. 2019. Does object recognition work for everyone?. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[13] Emily Denton and Timnit Gebru. 2020. Tutorial on Fairness, Accountability, Transparency and Ethics in Computer Vision.

[14] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. *COMPAS risk scales: Demonstrating accuracy equity and predictive parity.* Technical Report. Northpoint Inc.

[15] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning.*

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations.*

[17] Chris Dulhanty and Alexander Wong. 2019. Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[18] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research* 11, 5 (2010).

[19] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

[20] Christiane Fellbaum (Ed.). 1998. *WordNet: An electronic lexical database.* MIT press.

[21] Thomas B. Fitzpatrick. 1975. "Soleil et peau" [Sun and skin]. *Journal de Médecine Esthétique (in French)* 2 (1975), 33–34.

[22] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation* 80 (2016), 38.

[23] Gapminder. 2021. Dollar Street Dataset. https://www.gapminder.org/dollar-street. Retrieved: 2021-11-05.

[24] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems.*

[25] Paul Gölz, Anson Kahng, and Ariel D Procaccia. 2019. Paradoxes in fair machine learning. In *Conference on Neural Information Processing Systems.*

[26] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, and Piotr Bojanowski. 2021. Self-supervised Pretraining of Visual Features in the Wild. arXiv:2103.01988 [cs.CV]

[27] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *ACM Conference on Human Factors in Computing Systems.*

[28] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *ACM Conference on Fairness, Accountability, and Transparency.*

[29] Drew Harwell. 2019. *Federal study confirms racial bias of many facial-recognition systems, casts doubt on their expanding use.* https://www.washingtonpost.com/technology/2019/12/19/federal-study-confirms-racial-bias-many-facial-recognition-systems-casts-doubt-their-expanding-use/ Retrieved: 2021-11-05.

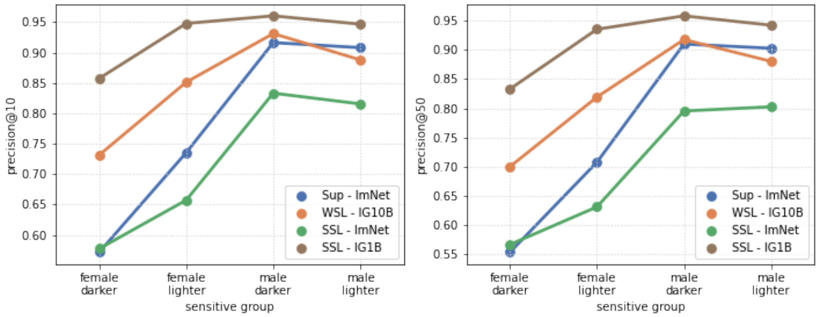


Figure 6: Same-group similarity search, precision@10 (left) and precision@50 (right) for the four main models compared.

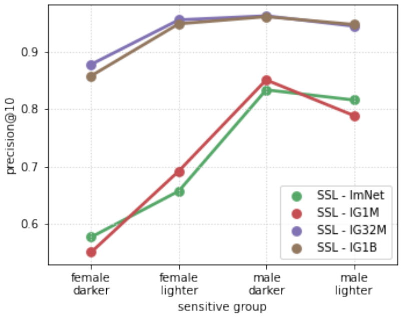


Figure 7: Similarity search, varying the data domain and size for SSL with RegNetY-128.

[30] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Towards Measuring Fairness in AI: the Casual Conversations Dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2021).

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[32] Radu Herbei and Marten H Wegkamp. 2006. Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* (2006).

[33] Minyoung Huh, Pulkit Agrawal, and Alexei A. Efros. 2016. What makes ImageNet good for transfer learning? arXiv:1608.08614 [cs.CV]

[34] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness. In *ACM Conference on Fairness, Accountability, and Transparency.*

[35] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *ACM Conference on Fairness, Accountability, and Transparency.*

[36] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning.*

[37] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision.* Springer, 67–84.

[38] Pratyusha Kalluri. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020), 169–169.

[39] Kimmo Karkkainen and Jungseock Joo. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision.*

[40] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 88 (nov 2018), 22 pages. https://doi.org/10.1145/3274357

[41] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv preprint arXiv:2005.04790* (2020).

[42] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv:2005.04790 [cs.AI]

[43] Davis E. King. 2009. Dlib-Ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* (Dec. 2009).

[44] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. 2015. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[45] Jon Kleinberg. 2018. Inherent Trade-Offs in Algorithmic Fairness. *SIGMETRICS Performance Evaluation Review* 46, 1 (jun 2018).

[46] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. 2020. Big Transfer (BiT): General Visual Representation Learning. In *European Conference on Computer Vision.*

[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Conference on Neural Information Processing Systems.*

[48] Bogdan Kulynych, David Madras, Smitha Milli, Inioluwa Deborah Raji, Angela Zhou, and Richard Zemel. 2020. Participatory Approaches to Machine Learning. In *International Conference on Machine Learning Workshops.*

[49] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4. *International Journal of Computer Vision* (2020).

[50] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[51] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation* 1, 4 (1989).

[52] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Conference on Neural Information Processing Systems.*

[53] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.*

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision.*

[55] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. 2018. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision.*

[56] Anthony Mccosker and Rowan Wilken. 2020. *Automating Vision: The Social Impact of the New Camera Consciousness.* Routledge.

[57] Anay Mehrotra and L. Elisa Celis. 2021. Mitigating Bias in Set Selection with Noisy Protected Attributes. In *ACM Conference on Fairness, Accountability, and Transparency.*

[58] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *ACM Conference on Fairness, Accountability, and Transparency.*

[59] Carina EA Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence* 3, 2 (2021), 104–110.

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

[61] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. 2020. Designing network design spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[62] Nitya Rajan. 2015. *Google Photo Algorithm Mislabels African Americans As 'Gorillas'.* https://www.huffingtonpost.co.uk/2015/06/30/goole-photo-algorithm-labels-african-american-as-gorillas_n_7696000.html/ Retrieved: 2021-12-01.

[63] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *AAAI/ACM Conference on AI, Ethics, and Society*.

[64] Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[65] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *AAAI/ACM Conference on AI, Ethics, and Society*.

[66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* (2015).

[67] Nicolas Schreuder and Evgenii Chzhen. 2021. Classification with abstention but without disparities. arXiv:2102.12258 [stat.ML]

[68] Candice Schumann, Caroline Rebecca Pantofaru, Susanna Ricco, Utsav Prabhu, and Vittorio Ferrari. 2021. A Step Toward More Inclusive People Annotations for Fairness. In *AAAI/ACM Conference on AI, Ethics, and Society*.

[69] Carsten Schwemmer, Carly Knight, Emily D. Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W. Lockhart. 2020. Diagnosing Gender Bias in Image Recognition Systems. *Socius* 6 (2020).

[70] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *NeurIPS 2017 workshop: Machine Learning for the Developing World*.

[71] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[72] Ryan Steed and Aylin Caliskan. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *ACM Conference on Fairness, Accountability, and Transparency*.

[73] Pierre Stock and Moustapha Cisse. 2018. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *European Conference on Computer Vision*.

[74] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[75] Alexander Tong. 2016. *Full ImageNet Taxonomy.* https://github.com/atong01/Imagenet-Tensorflow/blob/master/model/imagenet_synset_to_human_label_map.txt Retrieved: 2022-01-14.

[76] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. 2020. Clusterfit: Improving generalization of visual representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[77] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *ACM Conference on Fairness, Accountability, and Transparency*.

[78] Chicheng Zhang and Kamalika Chaudhuri. 2016. The extended Littlestone's dimension for learning with mistakes and abstentions. In *Conference on Learning Theory*.

[79] Zhifei Zhang, Yang Song, and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[80] Shoshana Zuboff. 2020. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.*

A DETAILED DESCRIPTIONS OF DATASETS

Casual Conversations (CC). [30] consists of 45, 186 videos of paid participants across five different cities in the US. Dataset provides *self-provided* age and gender labels in addition to annotated Fitzpatrick skin tone [21] and ambient lighting conditions. *Age* category varies from 18 to 85, *gender* is limited to 'male', 'female', 'other' and 'n/a' (not available, prefer not to say) and there is only a few videos for 'other' and 'n/a' gender categories. Nevertheless, CC is one of the few public datasets that has *self-identified* age and gender labels and hence we decided to assess our indicators also on this dataset. For the purpose of our indicators, we employ the CC dataset for model inference only and *do not* perform model training on this dataset. See Appendix B.1 for more details how to retrieve face crops.

OpenImages MIAP. [68] is a dataset designed to enable ML Fairness and constructed by providing additional annotations for a subset of OpenImages v6 [49] dataset. These annotations include bounding boxes of people and attribute labels for fairness such as *perceived* gender (predominantly masculine, predominantly feminine, unknown) presentation and *perceived* age range (young, middle, older, unknown).

For the purpose of our indicators, we use OpenImages MIAP *test* set for inference only. The *test* split contains 22, 590 images where each image has multiple bounded boxes. In order to focus on images of single people, we perform inference only on the bounding boxes with height and width >= 100 as otherwise the people are barely visible. We also filter out bounding boxes with gender or age of the unknown category as we observed they often are sketches, images of objects, or images where only the lower body is visible. The use of these images not critical in our fairness study, and we preferred keeping them out since it is unclear what label to expect from the classifier. Overall, we perform inference on 43, 523 effectively cropped images. We share more details on how we carefully apply the bounding box crop for inference in Appendix B.2.

UTKFaces. [79] is a large-scale face dataset that consists of 24, 108 face images each of which is annotated for *apparent* age, race and gender. The term *apparent* denotes that the labels are not *self-identified.* The dataset has almost balanced data between for male / female apparent genders with the downside of no data for non-binary gender. Images cover a variety of variations in pose, facial expression, illumination, occlusion and resolution. For the purpose of our proposed fairness indicator, we only use the apparent gender labels for which the dataset is also balanced.

Dollar Street. [23] is a collection of 16, 073 images of households captured all around the world by a group of photographers. The images capture 108 concepts from 289 households in 54 different countries across the world. The countries represent 4 different regions (The Americas, Europe, Asia, Africa). We manually mapped the 108 concepts to 94 classes in ImageNet [66] dataset [3]. The households have income levels (measured in USD) varying between 67\$ and $10k$. There are on average 53 unique images per household, with a maximum of 135 labeled objects per household.

[3] We release the full data information including the label mapping file to enable use of this indicator.

B EXPERIMENTAL DETAILS

B.1 Casual Conversations v1

For our experiments, we use the *mini* version of the dataset which consists of 2, 982 videos (two videos per participant with one *dark* and one *bright* lighting video when possible). Note that face crops are not available for 29 videos where DLIB [43] has failed to detect any faces. For each video, following [30], bounding boxes are detected on each frame of the video using DLIB [43] that are resized by a factor of 0.5 for faster computation and then upscaled by 2. In order to increase the area of face and background (hair and clothes) we enlarge the bounding boxes by a factor of 1.5 from the center on the aligned frames and resize the face crops to $256x256$. We take the middle frame of the video where there exists a face crop (when possible) and perform inference resulting on 2, 982 face images.

B.2 OpenImages

For OpenImages, we filter out bounding boxes with height or width smaller than 100 pixels, because we observed they correspond to barely visible people. We also filter out bounding boxes with gender or age of the unknown category as we observed they often are sketches, images of objects, or images where only the lower body is visible. The use of these images not critical in our fairness study, and we preferred keeping them out since it is unclear what label to expect from the classifier.

We then followed a two-step process to have bounding boxes fit the 224 × 224-input of our models:

(1) if the box's shape is too far from a square (we use the rule $\frac{\max(\text{height, width})}{\min(\text{height, width})} \geq 1.2$), we crop the bounding box further to a square from the top-left of the bounding box.
    That way, we obtain a square input with focus on the top of the bounding box, which usually is the head for images of people.
(2) we then resize the resulting image to 224x224

B.3 Adapting a visual system to predict labels

For the indicators proposed in Sec 3.1 and Sec. 3.2, visual systems need to be able to predict the labels and in particular labels in Dollar Street taxonomy for Geographical fairness indicator. There are 2 possible scenarios:

- **Visual system has label prediction capability (for Dollar Street taxonomy)**: If the dataset taxonomy of the dataset that the visual system is trained on, already predicts (Dollar Street) labels, such system can be used directly for inference on (Dollar Street) test images in *zero-shot* manner.
- **Visual systems don't have label prediction capability (and/or don't predict Dollar Street taxonomy)**: For some visual systems (such as self-supervised models) which are not inherently trained to predict any labels, the models can be adapted (for example by finetuning) to predict the labels by training on a *subset* of datasets like ImageNet [66] [4].

We also acknowledge that the caveat of using ImageNet as the transfer dataset can introduce the potential bias in the system but we note that this strategy can still enable comparisons of several visual systems conditioned that all the visual systems are adapted to Dollar Street taxonomy using the same training process (irrespective of if the models already have label prediction capability). We further note that our 3rd indicator allows to measure fairness purely from the raw model embeddings.

C LIMITATIONS IN MORE DETAIL

We note several limitations of our proposed indicators below, and some of them are the inevitable result of using currently available datasets. We note that these limitations would be easily overcome with more diverse and fairer datasets, which consist of all possible inclusive labels for all attributes.

- The proposed indicators provide a proof-of-concept for what could be the systematic assessment and evaluation of visual systems by utilizing the existing fairness datasets. These evaluations can enable comparison of models and measure how well they are calibrated (in particular, how do they impact marginalized populations). However, given a visual system, the choice of what indicator to measure depends on the context and this choice must be thoroughly assessed with proper stakeholder involvement so as to answer why those indicators are chosen, what kind of assumptions are embedded in this choice, and what specific questions do the system designers aim to answer [38, 48].
- We further note that these fairness indicators are complementary to the model and dataset documentation. These indicators ***do not*** replace proper documentation of dataset building and model reporting practices, but it can come together as a standard way of outlining a baseline comparison when one is developing or deploying new models. Furthermore, the proposed indicators can evolve/expand through time to include new types of visual systems, new fairness probes and new datasets as deemed appropriate.
- In our analysis of several visual systems, we only considered one model in each setup and didn't evaluate fairness if different seeds are used to initialize the models as this is beyond the scope of this work but we encourage using multiple seeds when probing the model fairness with the indicators.
- we focus on single-label prediction, mostly for convenience because classification is the most studied CV task in the literature. Even focusing only on classification, single-label prediction makes label ambiguity a problem – without all labels that could be considered correct for an image, error rates are unreliable.
- some limitations from the use of dataset:
    - the definitions of attributes differ across datasets which limits the possibility of cross-dataset tests (for instance, UTK faces doesn't have Fitzpatrick skintone scale in contrast to Casual Conversations. This limits the study to gender and age attributes in our Similarity search based indicator.
    - most dataset like UTK-Faces don't yet have data for non-binary genders which strictly limits the fairness probe for all social memberships.

[4] We release the information of exact subset of 20K subset images from ImageNet which capture the Dollar Street taxonomy and also subset labels that correspond to label associations in 3.1 indicator.

- most datasets only provide *perceived* or *apparent* labels for people's social membership which itself could be biased.
- lack of labels social membership labels (age, gender, skin-tone etc) puts constraints on the type of task we can address. For instance, publicly available Hate Speech datasets [41, 42] do not have labels for sensitive groups, and hence we can't analyze hate speech harms/biases of current visual systems towards specific groups this data.

D FURTHER CONTEXT AND RELATED WORK

Criticism of ImageNet. ImageNet [12, 66] has spurred immense developments and advances in Computer Vision over decades. However, recently Yang *et al.* [77] pointed out several reasons that ImageNet [12] might cause potential bias and therefore harm in the downstream models. Dulhanty and Wong [17] studied the demographics of people in ImageNet dataset by using computer vision models to predict the gender and age of people, and demonstrated that, e.g., males aged 15 to 29 make up the largest subgroup. Stock and Cisse [73] did not explicitly analyze the dataset but demonstrate that models trained on ImageNet exhibit misclassifications consistent with racial stereotypes. Steed *et al.* [72] further showed that unsupervised models trained on ImageNet can automatically learn racial, gender and intersectional biases from the way people portrayed in images that were curated from the web. Recently, effort has been made by Yang *et al.* [77] to reduce the bias in ImageNet dataset that resulted in removal of 2,702 `synsets` (out of 2,800 total) from `person subtree` in ImageNet dataset.

Fairness metrics for machine learning models. Most of the recent works utilize average performance indicators broken down by sensitive groups. The metrics themselves measure disparities between groups in terms of predicted positive value (`PPV`), True/False positive rates (`TPR`, `FPR`) [5], error rates [11] or risk difference [57]. There is no consensus nor general guidance for choosing the metric depending on context. Despite incompatibilities between metrics [7, 45] or otherwise counterarguments regarding the use of these aggregate performance measures [8, 25], there has been little debate on which metric to use for fairness audits in computer vision. This contrasts with the analysis of the COMPASS risk scores, which spurred an intense debate over metrics [14, 22, 50] (even though this debate is for the most part unresolved).

Centering fairness around the context of the task. Proper documentation of models and datasets are key priorities in developing new AI systems and should be carried out concurrently to benchmarks. Comprehensive approaches range from Model cards [58] aiming to standardize transparent model reporting, and data documentation [35] encouraging accountability at every stage of the data collection process. Raji *et al.* [65] pointed out five *ethical concerns* that should be taken into consideration while developing products for algorithmic auditing in order to prevent harms on protected groups. A common aspect of these developments is rooted in making the *context* of the task and the underlying cultural / social context as an important factor in developing transparent and accountable machine learning systems. Without such considerations, relying only on abstract notions of fairness measures fails to address the core problem of the developed system [34]. These challenges impact the researcher and the practitioner alike and the standardized protocols enabling systematic fairness assessment is a long overdue step forward for computer vision developments.

E ADDITIONAL LABEL ASSOCIATION RESULTS

In this section we provide additional label association results on Casual Conversations and OpenImages MIAP.

E.1 Results for different age groups.

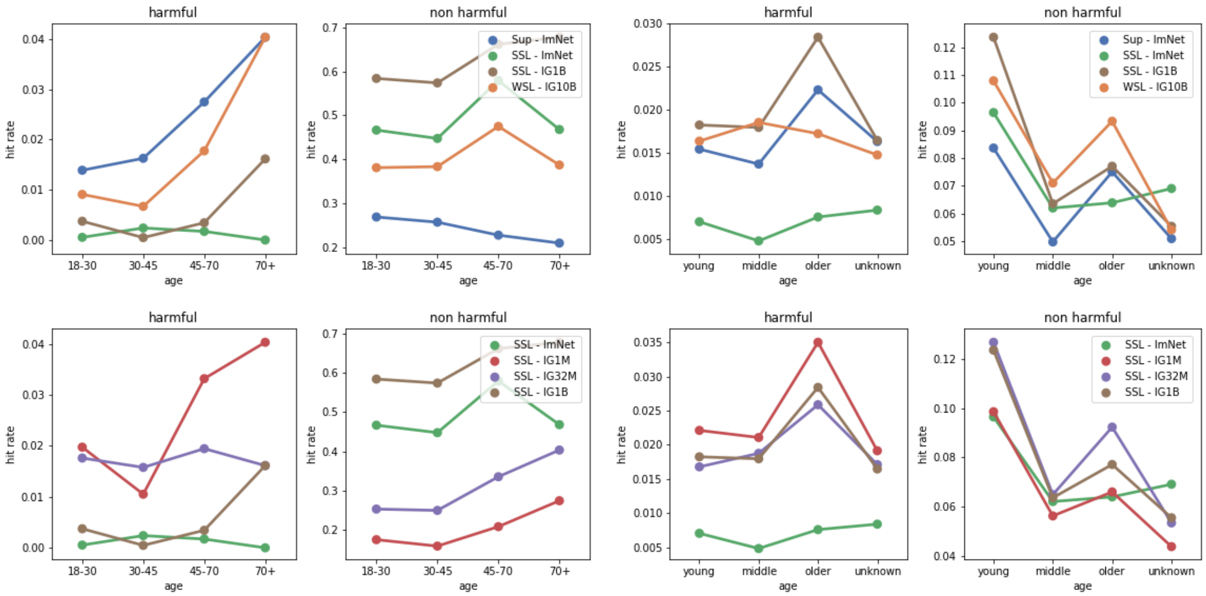
Following Section 3.1, all results are reported for the `RegNetY-128` model, considering predictions with a minimum confidence threshold $\tau = 0.1$.

Figure 8 depicts the impact of training paradigm, data size and domain on the potentially harmful label associations on both Casual Conversations and OpenImages MIAP when stratifying results based on (perceived) age. Results include `harmful` (`crime`, `non-human`) and `non-harmful` (`human`) predictions.

Effect of training paradigm. Figures 8a–8b (top) compare the effect of different training paradigms on Casual Conversations and OpenImages MIAP, respectively when stratifying results based on different age groups. As in the gender-skin tone stratification, we observe that trends appear to favor the SSL paradigm on Casual Conversations. However, we note that in this case, SSL-ImageNet reaches similar `harmful` hit rates as SSL-Instagram, except for the `70+` age group, where the `harmful` predictions are lower. The trend observed in the `non-harmful` predictions is the same as for the gender-skin tone stratification, with SSL-Instagram obtaining notably higher hit rates than SSL-ImageNet (possibly because Instagram data domain is more human-centric). In *OpenImages MIAP*, the analysis of the results also yields to similar observations as in the gender-skin tone classification. In particular, SSL-ImageNet continues to yield the lowest `harmful` hit rates, and models pretrained on Instagram continue among the top `harmful` label predictors. Notably, SSL-Instagram leads to the highest `harmful` hit rates. For `non-harmful` predictions, we observe a general trend that results in larger hit rate differences across different groups, no matter the training paradigm and data used.

Effect of data size and domain. Figures 3a–3b (bottom) show how data domain and data size affect the training of SSL models on Casual Conversations and OpenImages MIAP, respectively. The effect of data size when stratifying results based on age outlines the same trends as for the gender-skin tone stratification. Overall, it appears that increasing the data size from 1M to 1B Instagram images, drastically increases the `non-harmful` predictions and also lowers the `harmful` ones, especially on Casual Conversations. However, the SSL-ImageNet results in comparable or lower `harmful` prediction hit rates than the best Instagram-based model, further emphasizing that the data content might be more critical than the scale of the data to mitigate potentially harmful associations.

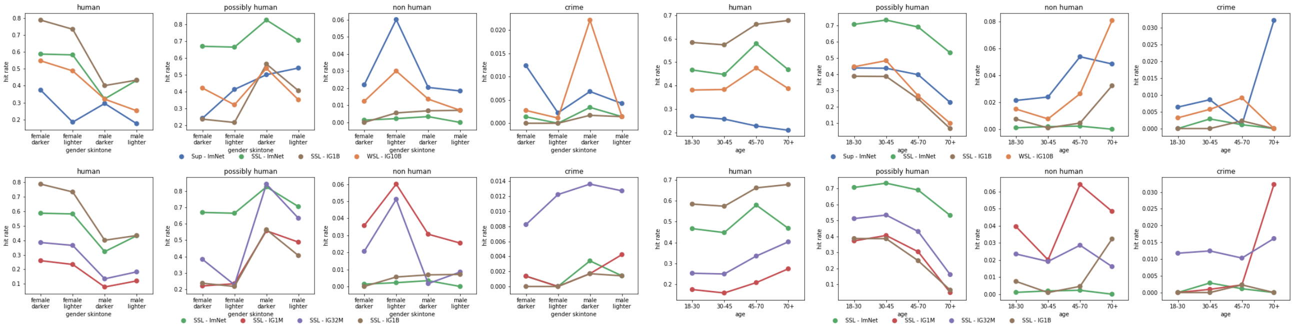
For completeness, Figure 9 presents label association results on Casual Conversations for all labels separately: `human`, `non-human`, `possibly-human`, and `crime`. Results are shown both for gender and skin tone, as well as age, and consider the effect of training



(a) Casual Conversations.

(b) OpenImages MIAP

Figure 8: Label association results stratified into groups based on age: (Top) Effect of training paradigm; (Bottom) Effect of data size and domain. For `harmful` labels, the lower the hit rate the better; conversely, for `non-harmful` labels the higher the hit rate the better.

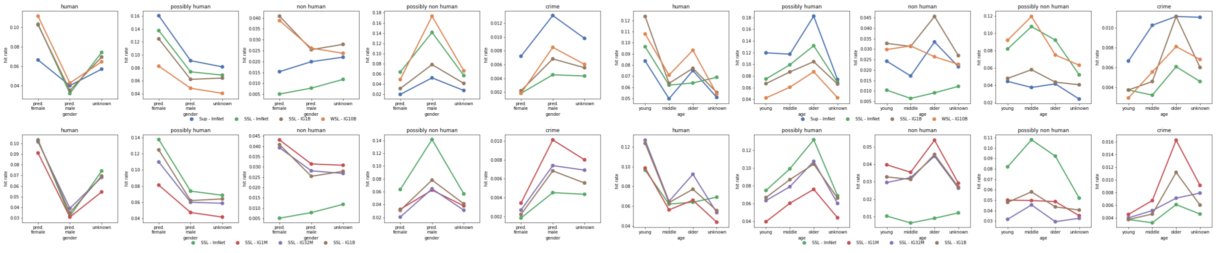


(a) Results based on gender and skin tone.

(b) Results based on age.

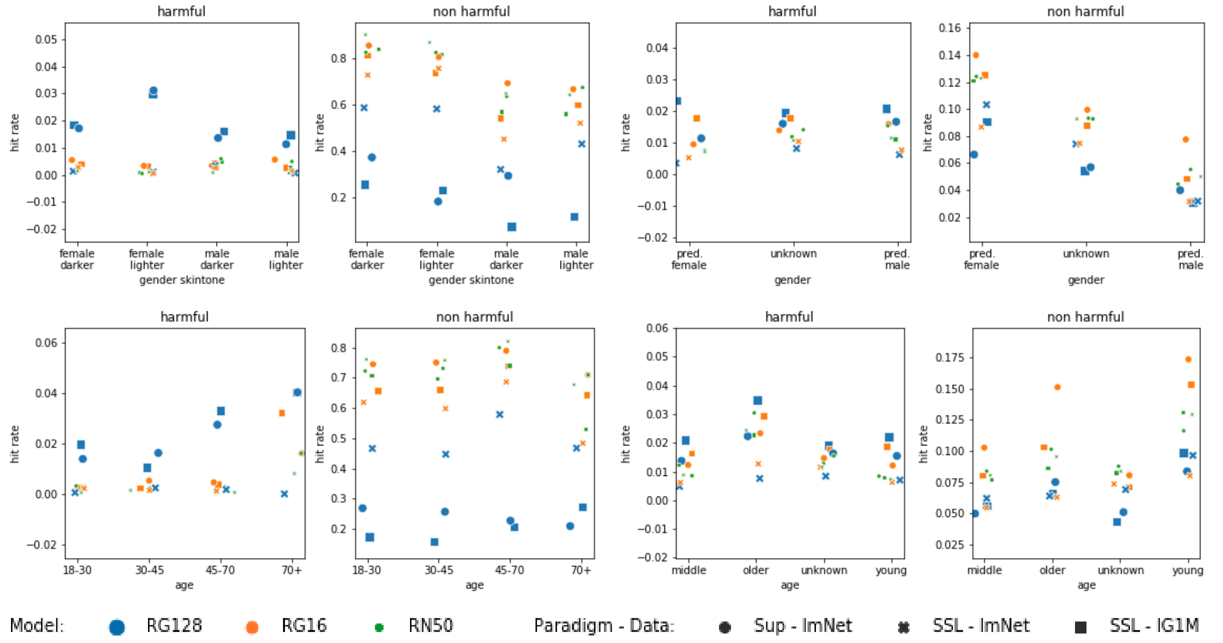
Figure 9: Casual Conversations extended label association results: (Top) Effect of training paradigm; (Mid) Effect of data size and domain. For `human` and `possibly-human`, the higher the better. For `crime` and `non-human`, the lower the better.

paradigm (top) and data size/domain (bottom). We note SSL leads to both lower `crime` and `non-human` hit rates. However, SSL-ImageNet resulting in overall lower `harmful` hit rates can be explained by its lower `non-human` prediction rates. SSL-Instagram results in lower `crime` prediction rates, and both its `crime` and `non-human` prediction rates can be decreased by appropriately increasing the dataset size (see IG1M vs IG1B results). Interestingly, SSL-ImageNet consistently exhibits among the highest `possibly-human` hit rates. Analogously, Figure 10 presents extended results on OpenImages MIAP. In this case, results are reported for `human`, `non-human`, `possibly-human`, `possibly-non-human`, and `crime`, and consider both perceived gender and age. When decoupling the `harmful` labels, we observe that the high `harmful` prediction rates of Instagram-pretrained models are driven by their frequent `non-human` predictions. However, when it comes to `crime`, Supervised-ImageNet leads to the highest prediction rates. In OpenImages MIAP, `possibly-non-human` predictions are more present in SSL models, whereas `possibly-human` predictions are more present in models pre-trained on ImageNet (no matter the paradigm).



(a) Results based on perceived gender. **(b) Results based on perceived age.**

Figure 10: OpenImages MIAP extended label association results: (Top) Effect of training paradigm; (Mid) Effect of data size and domain; (Bottom) Confidence of models.For `human` and `possibly-human`, the higher the better. For `crime`, `non-human` and `possibly-non-human`, the lower the better.



(a) Casual Conversations: (top) results based on gender and skin tone, (bottom): results based on age. **(b) OpenImages MIAP: (top) results based on perceived gender, (bottom): results based on perceived age.**

Figure 11: Effect of model capacity on label association results.

E.2 Effect of model capacity.

Finally, Figure 11 presents the impact of model capacity on label association results on both Casual Conversations and OpenImages MIAP. When considering Supervised-ImageNet models, we observe that increasing model capacity tends to increase the prediction of `harmful` labels across all groups considered in both Casual Conversations and OpenImages MIAP datasets. At the same time, the highest capacity model, `RegNetY-128`, exhibits the lowest `non-harmful` hit rates. This suggests that better fitting the ImageNet data with a supervised objective may lead to undesirable outcomes. When considering SSL-ImageNet models, increasing capacity maintains or decreases `harmful` hit rates across all groups and for both datasets considered. It is worth noting that these SSL models not only tend to exhibit among the lowest `harmful` hit rates but also tend to be more stable in their predictions across different subgroups. However, higher capacity SSL-ImageNet models often result in the lowest `non-harmful` hit rates. By contrast, when considering SSL-Instagram models, increasing model capacity often increases `harmful` predictions and leads to slightly higher discrepancies across different groups. In this case, similar to SSL-ImageNet, we observe that increasing model capacity often decreases or maintains the `non-harmful` predictions. However, as discussed in Section 3.1, these models can significantly benefit from additional Instagram data to mitigate potentially `harmful` label associations. It is worth noting that in the OpenImages MIAP, `non-harmful` prediction rates are overall significantly lower than in Casual Conversations.

F EFFECT OF MODEL CAPACITY: GEOGRAPHICAL DIVERSITY

We provide in Fig. 12 the results on the geographical fairness indicator depending on the model size, depending on the training paradigm (Supervised or SSL) and depending on data domain and size (ImageNet, or Instagram with 1*m*, 1*b*).

We observe that model capacity is critical when pre-training in a supervised fashion on ImageNet (leftmost column): from `ResNet-50` to `RegNetY-16` there is an improvement of over 15% (absolute) hit rate. `RegNetY-16` and `RegNetY-128` achieve the same performance on Supervised training. Interestingly, model capacity does not seem to have a large impact for SSL on 1*m* images, either on ImageNet or Instagram (middle left and middle right columns), and the performances dominate that of Supervised pre-training. The effect of model capacity becomes more visible when training on 1*b* images (rightmost column), where differences between `RegNetY-128` and `ResNet-50` are significant across all sensitive groups.

The conclusion is that the results coincide with those of other indicators when training at very large scale: larger models tend to fare better. Apart from that, the effect of model size on supervised training is interesting, yet seems specific to that indicator.

G EFFECT OF MODEL CAPACITY: SIMILARITY SEARCH

We present in this section the results obtained on the similarity search indicator by varying the model capacity. Fig. 13 presents the results in terms of precision@10 and precision@50 for `ResNet-50`, `RegNetY-16` and `RegNetY-128` for Supervised training on ImageNet (left most column), SSL on ImageNet (middle-left column), SSL on Instagram with 1*m* examples (middle-right) and SSL on Instagram with 1*b* examples.

As for all other indicators, increasing model capacity improves precision on SSL training on 1*b* examples, with `RegNetY-128` achieving the best precision values (and about 15% absolute improvement on `female darker`, the sensitive group where the precision is the lowest).

For other training paradigms/data sizes however, increasing model capacity does not help much. On Supervised pre-training, we see no effect of increasing the model capacity, while when using SSL on 1*m* examples (ImageNet or Instagram) `ResNet-50` tends to perform the best (on par with `RegNetY-128` on ImageNet, by far the best on Instagram with 1*m* examples). In conclusion, it seems that on this indicator, large-scale datasets are necessary for SSL with large models to shine. Yet, as noticed in Sec. 5.3, compared to supervised pre-training, the improvement of SSL on Instagram with 1*b* examples is substantial (comparing the leftmost and rightmost columns), with nearly 30% absolute improvement in precision@50 on `female darker` between supervised pre-training on ImageNet and `RegNetY-128` trained with SSL.

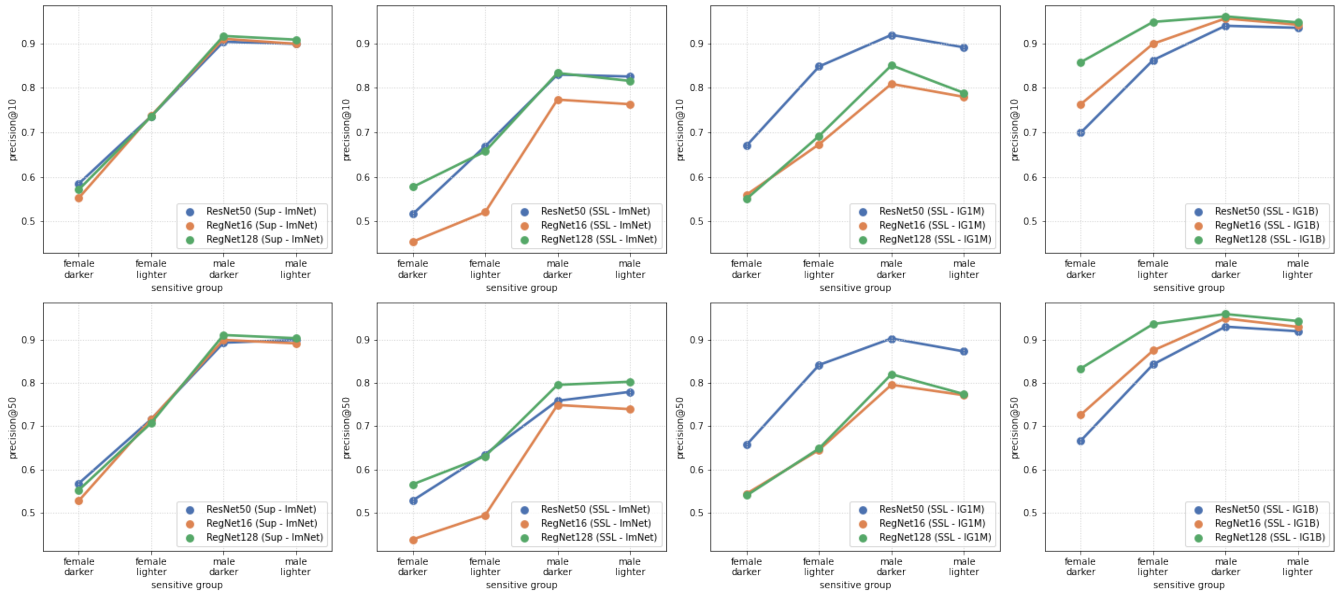


Figure 12: Geodiversity: effect of model capacity depending on the training paradigm and data size (columns), broken down by income bucket (top row) and region (bottom row).

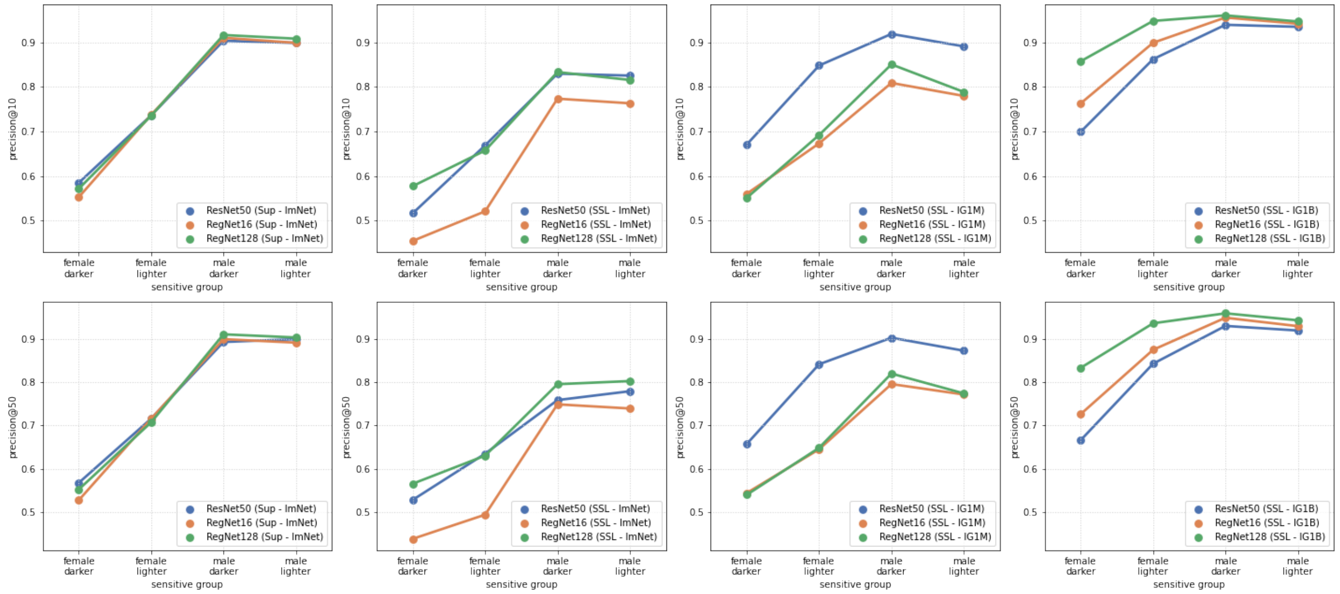


Figure 13: Similarity search indicator: precision@10 (top row) and precision@50 (bottom row) depending on model capacity, for different training paradigms, data domain and sizes (columns).