A VIEW FROM SOMEWHERE: HUMAN-CENTRIC FACE REPRESENTATIONS

Jerone T. A. Andrews*
Sony AI, Japan

Przemysław Joniak†
University of Tokyo, Japan

Alice Xiang
Sony AI, USA

ABSTRACT

Few datasets contain self-identified demographic information, inferring demographic information risks introducing additional biases, and collecting and storing data on sensitive attributes can carry legal risks. Besides, categorical demographic labels do not necessarily capture all the relevant dimensions of human diversity. We propose to implicitly learn a set of continuous face-varying dimensions, without ever asking an annotator to explicitly categorize a person. We uncover the dimensions by learning on A View From Somewhere (AVFS) dataset of 638,180 human judgments of face similarity. We demonstrate the utility of our learned embedding space for predicting face similarity judgments, collecting continuous face attribute values, attribute classification, and comparative dataset diversity auditing. Moreover, using a novel conditional framework, we show that an annotator's demographics influences the *importance* they place on different attributes when judging similarity, underscoring the *need* for diverse annotator groups to avoid biases. Data and code are available at https://github.com/SonyAI/a_view_from_somewhere.

1 INTRODUCTION

The canonical approach to evaluating human-centric image dataset diversity is based on demographic attributes labels. Many equate diversity with parity across the subgroup distributions (Kay et al., 2015; Schwemmer et al., 2020), presupposing access to demographically labeled samples. However, most datasets are web scraped, lacking ground-truth information about image subjects (Andrews et al., 2023). Moreover, data protection legislation considers demographic attributes to be personal information and limits their collection and use (Andrus et al., 2021; 2020).

Even when demographic labels are known, evaluating diversity based on subgroup counts fails to reflect the continuous nature of human phenotypic diversity (e.g., skin tone is often reduced to `light` vs. `dark`). Further, even within the same subpopulation, image subjects exhibit certain traits to a greater or lesser extent than others (Becerra-Riera et al., 2019; Carcagnì et al., 2015; Feliciano, 2016).

When labels are unknown, researchers typically choose certain attributes they consider to be relevant for human diversity and use human annotators to infer them (Karkkainen & Joo, 2021; Wang et al., 2019). Inferring labels, however, is difficult, especially for nebulous social constructs, e.g., race and gender (Hanna et al., 2020; Keyes, 2018) and can introduce additional biases (Freeman et al., 2011). The label taxonomies not only encode, reify, and propagate stereotypes beyond "their cultural context" (Khan & Fu, 2021), but also do not permit multi-group membership, resulting in the erasure of, e.g., multi-ethnic individuals (Robinson et al., 2020; Karkkainen & Joo, 2021). Significantly, discrepancies between inferred and self-identified attributes can induce psychological distress by invalidating an individual's self-image (Campbell & Troyer, 2007; Roth, 2016).

In this work, we avoid problematic semantic labels altogether and propose to learn a perceptual similarity function (i.e., model) aligned with human perception, measuring the similarity between two faces. As similarity is inversely connected to diversity (Leinster & Cobbold, 2012), our model naturally provides dataset users with a perceptual diversity measure. Our model, formulated under a novel conditional framework, not only learns to accurately predict human judgments of face similarity,

*Correspondence to `jerone.andrews@sony.com`.
†Work done while the author was an intern at Sony AI, Japan.

but also provides a human-interpretable decomposition of the dimensions used in the human-decision making process, as well as the importance distinct annotators place on each dimension. Underlying our model is A View From Somewhere (AVFS) dataset of 638,180 face similarity judgments over 4,921 faces. Each judgment corresponds to the odd-one-out (i.e., least similar) face in a triplet of faces and is accompanied by both the identifier and demographic attributes of the annotator who made the judgment. AVFS is made available under a Creative Commons license (CC-BY-NC-SA).

We demonstrate that (1) individual embedding dimensions learned by training on AVFS are related to concepts of gender, ethnicity, age, as well as face and hair morphology; (2) compared to face embeddings induced by learning on face identity and semantic face attribute datasets, our embeddings are highly correlated with the human mental representational space of faces; (3) annotators are influenced by their sociocultural backgrounds, underscoring the need for diverse annotator groups to mitigate bias; and (4) our model not only provides an effective similarity measure, but can also be used to collect continuous (as opposed to discrete) face attribute values for novel faces from annotators, binary attribute classification, and comparative dataset attribute disparity estimation.

2 Related Work

Face datasets. Most human face datasets are composed of semantically labeled images, created for the purposes of identity and attribute recognition (Karkkainen & Joo, 2021; Liu et al., 2015; Huang et al., 2008; Cao et al., 2018). When learning embeddings on such data, the implicit assumption is that semantic similarity is equivalent to visual similarity (Deselaers & Ferrari, 2011). However, many semantic categories (including human social categories) are functional (Rosch, 1975; Rothbart & Taylor, 1992), i.e., unconstrained by visual features such as shape, color, and material. Moreover, semantic labels only indicate the presence or absence of an attribute, as opposed to its magnitude, making it impossible to compare the similarity between same-labeled samples (Vemulapalli & Agarwala, 2019). Therefore, such data may not represent the best resource for learning a similarity function aligned with human visual perception. Unlike AVFS, existing face similarity datasets (Somai & Hancock, 2021; Sadovnik et al., 2018; McCauley et al., 2021; Vemulapalli & Agarwala, 2019) utilize synthetic imagery, have a limited number of judgments, are not publicly available, and/or wholly focus on lookalike or facial expression similarity.

Psychological embeddings. The human mind is conjectured by cognitive psychologists to have "a considerable investment in similarity" (Medin et al., 1993). When two entities are compared they mutually constrain the set of features that are activated or inferred in the human mind (Markman, 1996)—i.e., similarity is dynamic, where features are discovered and aligned based on what is being compared. Multidimensional scaling (MDS) is often used to learn psychological embeddings from human judgments of similarity (Zheng et al., 2019; Roads & Love, 2021; Dima et al., 2022; Josephs et al., 2021). As MDS approaches cannot embed images outside of the training set, researchers have used pretrained models as feature extractors (Sanders & Nosofsky, 2020; Peterson et al., 2018; Attarian et al., 2020), which can introduce implicit biases (Krishnakumar et al., 2021; Steed & Caliskan, 2021). In contrast, our method generalizes to novel face images and is trained end-to-end.

Annotator positionality. Attribute labeling by humans not only depends on the image subject being categorized, but also on the annotator's sociocultural background as well as extrinsic contextual cues (Segall et al., 1966; Balaresque & King, 2016; Hill, 2002; Roth, 2016; Freeman et al., 2011). Despite this, annotator positionality has only recently entered into discourse in computer vision (Chen & Joo, 2021; Zhao et al., 2021; Denton et al., 2021). Notably, "only five publications [from 113 surveyed] provided any [annotator] demographic information" (Scheuerman et al., 2021). In order to mitigate bias, one must first measure bias (Le Quy et al., 2022). To our knowledge, AVFS represents the first human-centric dataset, where each annotation is accompanied by both the identifier and demographic attributes of the annotator who generated it. In this work, we introduce a conditional framework that utilizes annotator identifiers, which increases predictive performance and reveals the importance distinct annotators place on different attributes when judging similarity.

3 A View From Somewhere Dataset

To learn a similarity function aligned with human perception, we collect, AVFS, a large-scale dataset of odd-one-out similarity judgments. An odd-one-out judgment corresponds to the least similar face

in a triplet of face images, representing a three alternative forced choice (3AFC) task. Refer to the Appendix for further details on AVFS.

Face image stimuli. For our proof of concept, 4,921 faces were sampled from the CC-BY licensed FFHQ (Karras et al., 2019) dataset. The subset was obtained by first splitting FFHQ into 56 partitions based on inferred intersectional group labels, and then randomly sampling from each partition with equal probability. Ethnicity was estimated using a FairFace model (Karkkainen & Joo, 2021); and binary gender expression and age group were obtained from FFHQ-Aging crowdsourced human annotations (Or-El et al., 2020). See the Appendix for more details. All faces are near-frontal with little to no eye occlusions and an apparent age > 19 years old.

3AFC similarity judgments. AVFS contains 638,180 quality-controlled triplets over 4,921 faces, representing 0.003% of all possible triplets. Annotators were presented with a triplet and instructed to: choose the person that looks least similar to the two other people (odd-one-out). To focus our proof of concept on intrinsic facial features, annotators were additionally instructed to ignore differences in pose, expression, lighting, accessories, background, and objects. Each AVFS triplet is labeled with a judgment, as well as the identifier of the annotator who made the judgment and their self-reported age, nationality, ancestry, and gender identity. As in previous non-facial odd-one-out datasets (Josephs et al., 2021; Hebart et al., 2022; Dima et al., 2022), there is a single judgment per triplet. Quality was controlled by excluding judgments from annotators who provided overly fast, deterministic, or incomplete responses. In total, 1,645 annotators contributed to AVFS via Amazon Mechanical Turk (AMT) and provided consent to use their study data. Compensation was 15 USD per hour.

3AFC task rationale. Let $\mathbf{x} \in \mathcal{X}$ and $\text{sim} : (\mathbf{x}_i, \mathbf{x}_j) \rightarrow \text{sim}(i, j) \in \mathbb{R}$ denote a face image and a similarity function, resp. Our motivation for collecting AVFS is fourfold. Most significantly, the odd-one-out task does not require an annotator to *explicitly* categorize people. Second, for a triplet $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$, repeatedly varying $\mathbf{x}_k$ permits the identification of the *relevant* dimensions that contribute to $\text{sim}(i, j)$. That is, wlog, $\mathbf{x}_k$ provides context for which $\text{sim}(i, j)$ is determined, making the task easier than *explicit* pairwise similarity tasks (i.e., "Is $\mathbf{x}_i$ similar to $\mathbf{x}_j$?"). This is because it is not always apparent to an annotator which dimensions are relevant when determining $\text{sim}(i, j)$, especially when $\mathbf{x}_i$ and $\mathbf{x}_j$ are perceptually different. As highlighted in several cognitive psychology works (Medin et al., 1993; Markman & Gentner, 2005; Goodman, 1972), context makes salient: context-related properties; and, the extent to which objects being compared share context-related properties. Third, there is no need to prespecify attribute lists hypothesized as relevant for comparison (e.g., "Is $\mathbf{x}_i$ *older* than $\mathbf{x}_j$?"). The odd-one-task implicitly encodes salient attributes that are used to determine similarity. Finally, compared to judgments for triplets composed of an anchor (i.e., reference point), $\mathbf{x}_a$, a positive, $\mathbf{x}_p$, and a negative, $\mathbf{x}_n$, odd-one-out judgments naturally provide more information. Odd-one-out triplets require an annotator to determine $\text{sim}(i, j)$, $\text{sim}(i, k)$, and $\text{sim}(j, k)$, whereas triplets with anchors only necessitate the evaluation of $\text{sim}(a, p)$ and $\text{sim}(a, n)$.

4 Conditional Perceptual Similarity Function

Zheng et al. (2019) developed an MDS approach for learning psychological embeddings from odd-one-out judgments, which have been shown to offer a *window* into the dimensions in the human mind of object categories (Hebart et al., 2020), human actions (Dima et al., 2022), and reachspace environments (Josephs et al., 2021). The approach is based on three assumptions. First, embeddings can be learned solely from odd-one-judgments, where representations are constrained to be continuous, non-negative, and sparse. Such properties support interpretability such that dimensions indicate both feature presence and feature magnitude. Second, odd-one-out judgments are a function of $\text{sim}(i, j)$, $\text{sim}(i, k)$, and $\text{sim}(j, k)$. Third, odd-one-out judgments are stochastic, where the probability of selecting x_k as the odd-one-out is $p(k) \propto \exp(\text{sim}(i, j))$.

Model of conditional decision-making. At a high-level, we want to apply Zheng et al. (2019)'s MDS approach to learn face embeddings. However, MDS cannot embed data outside of the training set, limiting its utility. Moreover, MDS pools all judgments, disregarding intra- and inter-annotator stochasticity. Therefore, we propose to learn a conditional convolutional neural network (CNN).

Let $\{(\{\mathbf{x}_{i\ell}, \mathbf{x}_{j\ell}, \mathbf{x}_{k\ell}\}, k\ell, a)\}_{\ell=1}^{n}$ denote a training set of n (triplet, judgment, annotator) tuples, where $a \in \mathcal{A}$. To simplify notation, we assume that judgments always correspond to index $k\ell$. Suppose $f : \mathbf{x} \mapsto f(\mathbf{x}) = \mathbf{w} \in \mathbb{R}^d$ is a CNN, parametrized by $\mathbf{\Theta}$, where $\mathbf{w} \in \mathbb{R}^d$ is an embedding

of $\mathbf{x}$. In contrast to Zheng et al. (2019), we model the probability of annotator a selecting $k\ell$ as $p(k\ell \mid a) \propto \exp(\text{sim}_a(i\ell, j\ell))$, where $\text{sim}_a(\cdot,\cdot)$ denotes the *internal* similarity function of a. Given two images $\mathbf{x}_i$ and $\mathbf{x}_j$, we define their similarity according to a as:

$$\text{sim}_a(i,j) = (\sigma(\boldsymbol{\phi}_a) \odot \text{ReLU}(\mathbf{w}_i))^\top \cdot (\sigma(\boldsymbol{\phi}_a) \odot \text{ReLU}(\mathbf{w}_j)), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function and $\sigma(\boldsymbol{\phi}_a) \in [0,1]^d$ is a *mask* associated with a. Each mask plays the role of an element-wise gating function, encoding the *importance* a places on each of the d embedding dimensions when determining similarity. Conditioning prediction on annotator identifiers induces subspaces that encode each annotator's notion of similarity, permitting us to study whether per-dimensional importance scores differ between annotators. Veit et al. (2017) employ a similar procedure to learn subspaces which encode different notions of similarity (e.g., font style, character type).

Conditional loss function. We denote by $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1^\top, \ldots, \boldsymbol{\phi}_{|\mathcal{A}|}^\top] \in \mathbb{R}^{d \times |\mathcal{A}|}$ a trainable weight matrix, where each column vector $\boldsymbol{\phi}_a^\top$ corresponds to annotator a's mask prior to applying $\sigma(\cdot)$. In our conditional framework, we jointly optimize $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ by minimizing:

$$-\sum_{\ell} \log\left[\hat{p}(k\ell \mid a)\right] + \alpha_1 \sum_i \|\text{ReLU}(\mathbf{w}_i)\|_1 + \alpha_2 \sum_{ij} \text{ReLU}(-\mathbf{w}_i)_j + \alpha_3 \sum_a \|\boldsymbol{\phi}_a\|_2^2, \tag{2}$$

where

$$\hat{p}(k\ell \mid a) = \frac{\exp(\text{sim}_a(i\ell, j\ell))}{\exp(\text{sim}_a(i\ell, j\ell)) + \exp(\text{sim}_a(i\ell, k\ell)) + \exp(\text{sim}_a(j\ell, k\ell))} \tag{3}$$

is the predicted probability that $k\ell$ is the odd-one-out conditioned on a. The first term in Eq. 2 encourages similar face pairs to result in large dot products. The second term (modulated by $\alpha_1 \in \mathbb{R}$) promotes sparsity. The third term (modulated by $\alpha_2 \in \mathbb{R}$) supports interpretability by penalizing negative values. The last term (modulated by $\alpha_3 \in \mathbb{R}$) penalizes large weights.

Our conditional decision-making framework is generally applicable to any task that involves mapping from inputs to decisions made by humans; it only requires record-keeping during data collection such that each judgment (or annotation) is associated with the annotator who generated it.

5 Experiments

We demonstrate the utility of our AVFS induced embedding spaces for predicting face similarity judgments, revealing annotator bias, collecting continuous face attribute values, attribute classification, and comparative dataset diversity auditing.

5.1 Experimental Setup

AVFS model training and implementation details. When $\alpha_3 = 0$ and $\sigma(\boldsymbol{\phi}_a) = [1, \ldots, 1]$ $(\forall a)$, Eq. 2 corresponds to the unconditional MDS objective proposed by Zheng et al. (2019). We refer to unconditional and conditional models trained on AVFS as AVFS-U and AVFS-C, resp., and conditional models whose masks are learned post hoc as AVFS-CPH. A AVFS-CPH model uses a fixed unconditional model (i.e., AVFS-U) to obtain face embeddings such that only $\boldsymbol{\Phi}$ is trainable.

AVFS models have ResNet18 (R18) (He et al., 2016) architectures and output 128-dimensional embeddings. We use the Adam (Kingma & Ba, 2014) optimizer with default parameters, reserving 10% of AVFS for validation. Based on grid search, we empirically set $\alpha_1 = 5 \times 10^{-5}$ and $\alpha_2 = 1 \times 10^{-2}$. For AVFS-CPH and AVFS-CPH, we additionally set $\alpha_3 = 1 \times 10^{-5}$. Across independent runs, we find that only a fraction of the 128 dimensions are needed and individual dimensions are reproducible. Post-optimization, we remove dimensions with maximal values close to zero. For AVFS-U, this results in 22 dimensions (61.9% validation accuracy). Note that we observe that 8/22 dimensions have a Pearson's r correlation > 0.9 with another dimension. Refer to the Appendix for further details on the models used in this paper.

Comparative embedding methods. Where appropriate, for comparison, we consider embeddings extracted from face identity verification, face attribute recognition, self-supervised, and object recognition models: (1–4) CASIA-WebFace (Yi et al., 2014) are face verification models trained with

Table 1: Predicting human similarity judgments. **(Same Stimuli)** Accuracy over 24,060 judgments and Spearman's r between the entropy in human- and model-generated triplet odd-one-out probabilities. **(Novel Stimuli)** Accuracy over 80,300 judgments and Spearman's r between the entropy in human- and model-generated similarity matrices. #Labels is an approximate of the number of labels collected to create a dataset accounting for consensus.

					Same Stimuli		Novel Stimuli	
Model	Method	#Images	#Labels	Arch.	Acc.	r	Acc.	r
ImageNet1K	Cross-entropy	1.3M	>1.3M	RG-32Gf	46.6	0.09	41.7	0.32
FairFace	Cross-entropy	62K	372K	R18	55.9	0.41	51.9	0.67
FairFace-L	Cross-entropy	62K	372K	R18	52.9	0.25	48.6	0.54
FairFace-BL	Cross-entropy	62K	372K	R18	51.7	0.15	47.7	0.54
CelebA	Cross-entropy	178K	7.1M	R18	52.1	0.25	48.9	0.56
CelebA-L	Cross-entropy	178K	7.1M	R18	53.2	0.15	49.6	0.59
CelebA-BL	Cross-entropy	178K	7.1M	R18	47.0	0.14	45.9	0.47
FFHQ	Cross-entropy	70K	140K	R18	50.5	0.16	47.3	0.49
FFHQ-L	Cross-entropy	70K	140K	R18	53.5	0.31	50.5	0.62
FFHQ-BL	Cross-entropy	70K	140K	R18	51.0	0.23	47.3	0.54
CASIA-WebFace	ArcFace	404K	404K	R18	51.6	0.29	46.1	0.40
CASIA-WebFace	Cross-entropy	404K	404K	R18	48.6	0.21	43.2	0.34
CASIA-WebFace	SphereFace	404K	404K	R18	48.3	0.25	43.8	0.35
CASIA-WebFace	CosFace	404K	404K	R18	44.0	0.12	40.8	0.23
AVFS-C	Conditional Equation (2)	5K	574K	R18	**67.4**	**0.68**	**61.7**	0.82
AVFS-CPH	Conditional Equation (2)	5K	574K	R18	66.5	**0.68**	61.4	0.82
AVFS-U	Unconditional Equation (2)	5K	574K	R18	62.0	0.65	57.5	**0.86**
AVFS-U	Unconditional Equation (2)	5K	287K	R18	61.3	0.61	55.8	0.81
AVFS-U	Unconditional Equation (2)	5K	144K	R18	60.6	0.56	55.3	0.80
AVFS-U	Unconditional Equation (2)	5K	72K	R18	58.6	0.47	55.0	0.79
AVFS-Triplet	Triplet margin with distance swap (Balntas et al., 2016)	5K	574K	R18	60.2	0.46	52.8	0.64
ImageNet1K	SwAV	1.3M	0	RN50-w5	43.8	0.08	41.0	0.30
IG-1B	SwAV	1B	0	RG-32Gf	47.2	0.18	44.7	0.45
IG-1B	SwAV	1B	0	RG-64Gf	48.1	0.16	44.4	0.45
IG-1B	SwAV	1B	0	RG-128Gf	46.8	0.15	42.8	0.40
IG-1B	SwAV	1B	0	RG-256Gf	47.8	0.17	43.1	0.41
PASS	MoCo-v2	1.3M	0	R50	42.3	0.09	40.5	0.27
PASS	SwAV	1.3M	0	R50	42.4	0.12	40.4	0.27
PASS	DINO	1.3M	0	ViTS-16	43.2	0.10	41.8	0.32

a cross-entropy, ArcFace (Deng et al., 2019), SphereFace (Liu et al., 2017), and CosFace (Wang et al., 2018) loss on images from 11K identities; (5) CelebA (Liu et al., 2015) is a face attribute model trained to predict 40 binary attribute labels (e.g., `Pale Skin`, `Young`, `Male`); (6–7) FairFace (Karkkainen & Joo, 2021) and FFHQ are face attribute models trained to predict gender, age, and ethnicity; (8–11) IG-1B (Goyal et al., 2022) utilize the self-supervised SwAV (Caron et al., 2020) framework trained on uncurated Instagram images, containing millions of images of humans; (12–14) PASS (Asano et al., 2021) utilize self-supervised SwAV, MoCo-v2 (Chen et al., 2020), and DINO (Caron et al., 2021) frameworks trained on images without people; (15–16) ImageNet1K (Russakovsky et al., 2015) are a object recognition model and a self-supervised SwAV framework trained on 1.3M images with 17% of the images containing at least one human face (Yang et al., 2022).

Methods (1–7) have R18 architectures and output 128-dimensional embeddings, whereas (8–16) correspond to official implementations that vary wrt architecture and output dimensionality. For (5–7), we additionally consider unnormalized class logit embeddings (e.g., denoted CelebA-L), as well as logit embeddings converted to binary vectors based on class predictions (e.g., denoted CelebA-BL). All baseline embeddings are normalized to unit-length, where the dot product of two embeddings determines their similarity.

5.2 Same Stimuli

Predicting human similarity judgments. The utility of an embedding method may be determined by measuring whether similar stimuli are closer together in feature space than dissimilar stimuli. We first analyze whether our model results in embeddings of the stimulus set such that we are able to predict human judgments not observed during learning. Using images from the stimulus set of 4,921 faces, we generate 1,000 novel triplets and collect 22–25 unique judgments on AMT per triplet (24,060 judgments). In addition to odd-one-out triplet predictive accuracy, we report Spearman's r correlation between the entropy in human- and model-generated triplet odd-one-out probabilities. Human-generated odd-one-out probabilities are of the form: $(n_i, n_j, n_k)/n$, where, wlog, n_k/n corresponds to the fraction of n odd-one-out votes for k.

As we have 22–25 judgments per triplet, we can reliably estimate odd-one-out probabilities. The Bayes optimal classifier accuracy corresponds to the best possible accuracy any model could achieve given the stochasticity in the human judgments. The classifier makes the most probable prediction, i.e., the majority judgment. Thus, its accuracy is equal to the mean majority judgment probability over the 1,000 triplets, corresponding to 65.5$\pm$1% (95% CI).

Figure 1: For a subset of the 22 AVFS-U dimensions, we show: **(a)** the highest scoring stimulus set images; **(b)** the highest frequency topic labels generated by human annotators; and **(c)** dimension labels generated using CelebA and FairFace face attribute recognition models.

Results are shown in Table 1, evidencing that AVFS trained models outperform the baselines even when learning is performed on a fraction of the available judgments (e.g., 72K judgments). Table 1 shows three interesting results. First, the uncertainty in human- and AVFS model-generated triplet odd-one-out probabilities are highly correlated, underscoring the human-like ability of our model. Second, our conditional models have increased performance over their unconditional counterparts, showing that the learned annotator-specific masks generalize. Moreover, our conditional models attain a predictive accuracy at or above the upper bound of the Bayes optimal classifier. Finally, as posited in Section 3, transforming AVFS into a dataset of standard triplets constraints with anchors results in lower performance due to a reduction in information.

Are the learned dimensions human-interpretable? Since we aim to replace the *explicit* collection of problematic categorical labels, we evaluate whether the individual dimensions are human-interpretable through a qualitative dimension labeling task (Hebart et al., 2020; Josephs et al., 2021). In what follows, let x_i and $\mathrm{ReLU}(w_i)$ denote an image from the stimulus set of 4,921 faces and its corresponding 22-dimensional embedding, resp. Further, let $\%\mathrm{tile}_j^q$ denote the q-th percentile of the dimension j embedding values. We define grid_j as a 5×100 grid of faces. Column $q \in [100, \ldots, 1]$ of grid_j corresponds to faces from the stimulus set of 4,921 faces with the top 5 highest dimension j embedding values, satisfying $\%\mathrm{tile}_j^{q-1} \leq \mathrm{ReLU}(w)_j < \%\mathrm{tile}_j^q$. Thus, from left (high) to right (low), grid_j shows example faces from each percentile for dimension j. We task annotators with writing 1–3 visual characteristics that describe grid_j. For each grid_j, we collect 25–62 labels. As the task is open-ended, we manually convert the labels into 35 broad topics.

Figure 1 provides evidence of the interpretability for a random subset of AVFS-U dimensions. There is clear relationship between the dimension topics obtained from annotator descriptions and dimension labels generated using CelebA and FairFace models. Across the 22 dimensions, we note the materialization of individual dimensions coinciding with commonly defined demographic groups, i.e., `Male`, `Female`, `Black`, `White`, `East Asian`, `South Asian`, and `Elderly`. In addition, separate dimensions surfaced for face and hair morphology, i.e., `Wide Face`, `Long Face`, `Smiling Expression`, `Neutral Expression`, `Balding`, `Facial Hair`, and `Dyed Hair`. This is achieved without ever asking an annotator to explicitly categorize a person.

5.3 Novel Stimuli

Predicting human similarity judgments. Next, we analyze whether our model transfers to novel stimuli. To do so, we sample 56 novel face images from FFHQ not contained in the stimulus set. We then generate all $\binom{56}{3}$ possible triplets and collect 2–3 unique judgments on AMT per triplet (80,300 judgments). In addition to odd-one-out predictive accuracy, we report Spearman's r between the strictly upper triangular model- and human-generated similarity matrices. Entry (i, j) in the human-generated similarity matrix corresponds to the fraction of triplets containing (i, j), where neither was judged as the odd-one-out. Entry (i, j) in a model-generated similarity matrix corresponds to the mean $\hat{p}(i, j)$ over all triplets containing (i, j).

Results are again shown in Table 1 and largely follow the same trend to what we observe in the same stimuli setting. The results underscore that AVFS models generalize to arbitrary, novel face image stimuli, in particular, based on the correlation tests, the AVFS induced embedding space accurately reflects the human mental representational space of faces. This is significant as it shows that our model provides an effective similarity measure well-aligned with human perception.

Table 2: **(a)** Average AUC of linear SVM trained to discriminate between AVFS-CPH annotator masks from different demographic groups. **(b)** Spearman's r between CFD face attribute typicality ratings and model-generated attribute values.

(a)

Annotator Attribute	Groups	#Masks	AUC
Age Group	30-39 / 40-49	393 / 121	0.59 ± 0.05
Gender Identity	Male / Female	523 / 473	0.65 ± 0.05
Nationality	America / India	530 / 204	**0.86 ± 0.03**
Regional Ancestry	Europe / Asia	407 / 243	**0.86 ± 0.03**
Subregional Ancestry	West Europe / South Asia	173 / 107	**0.88 ± 0.05**

(b)

	Attribute				
Model	Masculine	Feminine	East Asian	Black	White
AVFS-U	0.813	0.840	0.747	**0.683**	0.644
CelebA	**0.858**	**0.855**	—	—	0.123
FairFace	0.836	0.827	**0.806**	0.653	0.645
FFHQ	0.812	0.829	0.794	0.587	**0.688**

Number of dimensions required to represent a face. To quantify the number of dimensions required to represent a face while preserving performance, we follow Hebart et al. (2020)'s dimension-elimination approach. We iteratively zero out the lowest value per face embedding until a single nonzero dimension remains. As this is done per embedding, the same dimension is not necessarily zeroed out from all embeddings during the same iteration. To obtain 95–99% of the predictive accuracy 6–13 dimensions are required, whereas to explain 95–99% of the variance in the similarity matrix we require 15–22 dimensions. This shows that (1) humans utilize a larger number of dimensions to represent the global similarity structure of faces than for determining individual odd-one-out judgments; and (2) similarity judgments are dynamic and context-dependent.

Are annotators interchangeable? Conditioning prediction on annotator identifiers provides the best predictive accuracy, evidencing that knowledge of the annotator determining similarity assists in informing the outcome. However, annotators are often framed as interchangeable (Malevé, 2020; Chancellor et al., 2019). To test the validity of this assumption, we randomly swap the annotator associated with each judgment and then recompute the predictive accuracy using the AVFS-CPH model. Repeating this process 100 times results in a performance drop from 61.7% to 52.8% ± 0.02% (95% CI) on average. This shows that annotator subspaces, and hence annotators, are not interchangeable.

Are annotators influenced by their sociocultural background? An interesting question relates to whether an annotator's sociocultural background influences their decision making. To evaluate this, we create datasets $\{(\sigma(v_a), y)\}$, where $\sigma(v_a)$ and $y \in \mathcal{Y}$ are annotator a's learned mask and self-identified demographic attribute, resp. For a particular annotator attribute (e.g., nationality), we limit the dataset to annotators who contributed ≥ 200 judgments and belong to one of the two largest groups wrt a self-identified demographic attribute. Using 10-fold cross validation, we train linear SVMs (Hearst et al., 1998) with balanced class weights to predict y from $\sigma(v_a)$. Table 2a shows the average AUC for each attribute. Most significantly, none of the AUC confidence intervals include chance performance. The linear SVMs are able to discriminate between binary groups wrt nationality, regional ancestry, and subregional ancestry with high probability (86–88%).

Continuous attribute value collection via the learned dimensions. Let μ_j^q denote the mean value of AVFS-U dimension j embedding values of *all* 4,921 faces from the stimulus set, satisfying $\%\text{tile}_j^{q-1} \leq \text{ReLU}(w)_j < \%\text{tile}_j^q$. For the *dimension rating task*, annotators must place novel faces (e.g., x) above a single column q of grid_j. Placement is based on the similarity between x and the faces contained in each column q of grid_j. Annotators are not primed with the meaning of grid_j. If x is accurately placed on grid_j, then this indicates that the ordering of dimension j is visually coherent. To test this, we sample 20 novel faces from FFHQ, i.e., the faces are not contained in the stimulus set. $\forall(x, \text{grid}_j)$, we collect 20 unique judgments (8,800 judgments). From the judgments, for each x, we create 22-dimensional human-generated embeddings of the form: $[\frac{1}{20}\sum_{n=1}^{20}\mu_1^{a_n}, \ldots, \frac{1}{20}\sum_{n=1}^{20}\mu_{22}^{a_n}]$, where $a_n \in [100, \ldots, 1]$ denotes the n-th annotator's column choice q. Next, we generate all $\binom{20}{3}$ possible triplets to create a human-generated similarity matrix $M \in \mathbb{R}^{20\times20}$. Entry (i, j) in M corresponds to the mean $\hat{p}(i, j)$ over all triplets containing (i, j) using the human-generated embeddings. We create a model-generated matrix in the same way, but this time using the model-generated embeddings.

Spearman's r correlation between the strictly upper triangular model- and human-generated similarity matrices is 0.83 and 0.86 for AVFS-U and AVFS-C-PH, resp. This shows that (1) dimensional values correspond to the feature magnitude; and (2) image grids can be used to *directly* collect continuous

Table 3: Results for **(a)** semantic binary-valued attribute classification AUC; and **(b)** comparative dataset attribute disparity estimation. (Note that $\star$ denotes that the ground-truth attribute value is self-reported by an image subject, whereas gray values indicate that a model was trained on the same data it is tested on.)

(a)

Face Data	Attribute	Model (AUC) AVFS-U	CelebA	FairFace	FFHQ
CasCon	$>$ 70 y.o.*	0.800	0.936	0.959	**0.962**
FFHQ	$>$ 70 y.o.	0.905	0.959	**0.971**	0.980
CFD	Male*	0.991	0.997	0.996	**0.998**
CasCon	Male*	0.971	0.986	**0.990**	0.981
CelebA	Male	0.990	0.999	0.994	**0.995**
COCO	Male	0.893	0.926	**0.963**	0.958
MIAP	Male	0.924	0.942	**0.945**	0.938
FFHQ	Male	0.933	0.959	**0.988**	0.996
CelebA	Smiling	**0.895**	0.982	—	—
CFD	Happy	0.969	**0.992**	—	—
CFD	Neutral	**0.731**	—	—	—
CFD	East Asian*	**0.969**	—	0.955	0.938
CFD	Black*	**0.992**	—	**0.992**	0.988
CFD	White*	0.972	0.889	**0.990**	0.988
CFD	Indian*	**0.960**	—	0.848	0.883
CasCon	Light skin	0.930	0.830	**0.965**	0.960
COCO	Light skin	0.889	0.771	0.927	**0.931**
CelebA	Balding	**0.963**	0.995	—	—

(b)

Dataset	Attribute	Model (Disparity Δ / Spearman's r) AVFS-U	CelebA	FairFace	FFHQ
CasCon	$>$ 70 y.o.*	0.22 / 0.96	**0.06 / 0.99**	**0.06** / 0.95	0.08 / **0.99**
FFHQ	$>$ 70 y.o.	0.16 / 0.95	0.10 / 0.89	**0.02 / 0.99**	0.02 / 1.00
CFD	Male*	**0.00 / 1.00**	0.08 / 0.99	0.02 / **1.00**	0.02 / **1.00**
CasCon	Male*	0.06 / 0.97	**0.02** / 0.99	**0.02 / 1.00**	0.04 / 0.99
CelebA	Male	0.10 / 0.96	0.04 / 1.00	0.12 / 0.96	**0.04 / 0.98**
COCO	Male	0.06 / 0.98	0.06 / 0.97	0.06 / 0.99	**0.00 / 1.00**
MIAP	Male	**0.04** / 0.99	**0.04 / 1.00**	**0.04** / 0.97	0.06 / 0.99
FFHQ	Male	0.14 / 0.83	0.14 / 0.93	**0.08 / 1.00**	0.04 / 1.00
CelebA	Smiling	**0.30 / 0.78**	0.14 / 0.97	—	—
CFD	Happy	0.12 / **0.99**	0.06 / **0.95**	—	—
CFD	East Asian*	0.06 / **1.00**	—	**0.02** / 0.98	**0.02 / 1.00**
CFD	Black*	0.10 / 0.98	—	0.14 / 0.97	**0.00 / 1.00**
CFD	White*	0.04 / **1.00**	**0.02** / 0.99	0.04 / 0.97	0.04 / 0.99
CFD	Indian*	**0.08 / 1.00**	—	0.30 / 0.73	**0.08** / 0.87
CasCon	Light skin	**0.06** / 0.94	0.3 / 0.65	0.18 / 0.85	**0.06 / 0.99**
COCO	Light skin	0.22 / 0.89	0.96 / 0.02	0.24 / 0.87	**0.02 / 1.00**
CelebA	Balding	**0.10 / 0.98**	0.06 / 1.00	—	—

attribute values for faces, sidestepping the limits of categorical definitions (Keyes, 2018; Benthall & Haynes, 2019; Khan & Fu, 2021). Note that the use of an image grid is not restricted to any particular embedding method, as long as the dimension is human-interpretable.

Correlation between feature magnitude and prototypicality. In cognitive science, the prototypicality of an entity corresponds to the extent to which it "belongs" to a conceptual category (Rosch, 1973). Having shown that a proportion of our model's dimensions are related to conceptual social categories, we now evaluate whether the value of a face along a dimension corresponds to its typicality. We utilize face images labeled with prototypicality ratings from the Chicago Face Database (CFD) (Ma et al., 2015; Lakshmi et al., 2021; Ma et al., 2021). The ratings (obtained from human annotators) correspond to the average prototypicality of a face wrt a race category from one (less typical) to five (very typical), considering skin color, hair, eyes, nose, cheeks, lips, and other physical features. For gender expression, ratings correspond to the typicality of the face relative to others of the same race and gender in the US from one (not at all typical) to seven (extremely typical). For each conceptual category, we extract a face's dimensional value or unnormalized attribute logit from a relevant AVFS-U dimension or attribute recognition model's classification layer, resp.

Table 2b shows that relevant AVFS-U dimensions are positively correlated with the typicality ratings according to Spearman's r. This (1) adds to the evidence that dimensional values correspond to feature magnitude; and (2) shows category typicality, at least for the investigated social category concepts, manifests in the learned dimensions from learning to predict human similarity judgments.

Semantic classification without semantically labeled training data. In light of the human-interpretability of our model's dimensions and results confirming that dimensional values correspond to the degree to which an attribute manifests in a face, we compare the AVFS-U dimensions with face attribute recognition models on the task of semantic binary-valued attribute classification. For face data, we use the following datasets: COCO (Lin et al., 2014), OpenImages MIAP (MIAP) (Schumann et al., 2021), CFD, FFHQ, CelebA (Liu et al., 2015), and Casual Conversations (CasCon) (Hazirbas et al., 2021). Continuous face attribute values are extracted in the same manner as in the earlier prototypicality experiment.

As shown in Table 3a, AVFS-U dimensions are competitive with face attribute recognition models, even in challenging unconstrained settings (i.e., COCO and MIAP). Training on AVFS not only results in a human-interpretable decomposition of the dimensions used by humans when determining face similarity, but also dimensions that can individually serve as semantic attribute classifiers.

Comparative dataset attribute disparity estimation. Finally, inspired by biodiversity measures (Leinster & Cobbold, 2012), we propose to use our learned dimensions for comparative dataset attribute disparity estimation. Concretely, given a set of candidate datasets $\{\mathcal{D}_k\}_{k=1}^n$ and a binary-valued attribute y, we aim to find $\mathcal{D}^\star \in \{\mathcal{D}_k\}_{k=1}^n$ with the smallest attribute disparity. Let $q_k \in [0, 0.01, \ldots, 0.99, 1]$ and $1 - q_k$ denote the proportion of images $\mathbf{x} \in \mathcal{D}_k$ labeled $y = 0$ and $y = 1$, resp. We define the similarity between $(\mathbf{x}_i, \mathbf{x}_j) \sim \mathcal{D}$ as $\text{abs}(\hat{y}_i - \hat{y}_j)^{-1}$, where $\hat{y} \in \mathbb{R}$ is a

predicted continuous attribute value for y given $\mathbf{x}$. In biodiversity terms, the average ordinariness of faces in a the set of samples, $\mathcal{D}$, is $\propto \sum_i \sum_j \text{abs}(\hat{y}_i - \hat{y}_j)^{-1}$ $(\forall i)$. This quantity is large when most faces in $\mathcal{D}$ are concentrated into a few very similar faces. As concentration is inversely connected to diversity, we can interpret the average dissimilarity, i.e., $|\mathcal{D}|^{-2} \sum_i \sum_j \text{abs}(\hat{y}_i - \hat{y}_j)$, as a diversity score.

For face data, we use COCO, MIAP, CFD, FFHQ, CelebA, and CasCon. To determine $\mathcal{D}^\star$, wrt an attribute y, we extract a face's dimensional value or unnormalized attribute logit from a relevant AVFS-U dimension or attribute recognition model's classification layer, resp. We bootstrap each $\mathcal{D}_k$ 100 times, where $|\mathcal{D}_k| = 100$ $(\forall k)$. We report (1) the attribute disparity in $\mathcal{D}^\star$, i.e., $\Delta = \text{abs}(2q_k - 1)$; and (2) Spearman's r correlation between the k bootstrapped average diversity scores and the k ground-truth disparity scores. Results are shown in Table 3b and confirm that AVFS-U dimensions are competitive with the baselines, despite not being learned on semantically labeled data. Spearman's r highlights that AVFS-U disparity scores are highly correlated with ground truth disparity based on labels.

6 Discussion and Conclusion

We proposed a method for *implicitly* learning continuous face-varying dimensions, without ever asking an annotator to explicitly categorize a person. We uncovered the face embedding space by learning on a novel dataset of human judgments of face similarity (AVFS). We showed that the individual dimensions are human-interpretable and related to concepts of gender, race, age, as well as face and hair morphology categories. We demonstrated the utility of our learned embedding space for predicting face similarity judgments, collecting continuous face attribute values, binary-attribute classification, and comparative dataset attribute diversity auditing. Moreover, using a novel conditional framework, we showed that an annotator's demographics influences the *importance* they place on different attributes when judging similarity, underscoring the *need* for diverse annotator groups to avoid biases. As our conditional decision-making framework is generally applicable to any task that involves mapping from inputs to decisions made by humans, it would be interesting to study the behavior of annotators in other visual tasks.

Our work is not without its limitations. First, we assume that the stimuli set is sufficiently diverse. Therefore, our current proposal is limited to the proof of concept that our approach can only uncover factors that vary in the data and are salient to human perception of face similarity. For instance, if skin color does not vary among data instances, then skin color cannot possibly influence human judgments, rendering it impossible to learn skin color as a face dimension. Second, the number of (near) nonzero dimensions depends on the $L1$ sparsity penalty, which must be carefully chosen. Too high of a penalty will result in the entanglement of distinct factors of variation, whereas too low of a penalty will result in repeated dimensions. We erred on the side of caution (lower penalty) to avoid merging distinct factors, which would reduce interpretability. Third, we found that participants did not always follow our instructions in full. Dimensions emerged corresponding to facial expression, which participants were instructed to ignore. Finally, our participant pool was demographically imbalanced, which is an unfortunate consequence of using AMT.

Despite the limitations, we have shown that similarity judgments are a valuable means by which to reveal the human mental representations of faces. Of particular note is the fact this was achieved by learning on 0.003% of all possible triplets from 4,921 faces. We hope that our work inspires others to pursue unorthodox tasks for learning the dimensions of human diversity, which do not require annotators to *explicitly* categorize people. Moreover, we aim to increase discourse on annotator positionality. As the philosopher Thomas Nagel suggested, it is impossible to take a "view from *nowhere*" (Nagel, 1989). Therefore, we need to make datasets more inclusive by integrating a diverse set of perspectives from their inception.

References

Jerone TA Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. *arXiv preprint arXiv:2302.03629*, 2023.

McKane Andrus, Elena Spitzer, and Alice Xiang. Working to address algorithmic bias? don't overlook the role of demographic data. *Partnership on AI. Retrieved from https://www. partnershiponai. org/demographic-data*, 2020.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 249–260, 2021.

Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *arXiv preprint arXiv:2109.13228*, 2021.

Maria Attarian, Brett D Roads, and Michael C Mozer. Transforming neural network visual representations to predict human judgments of similarity. *arXiv preprint arXiv:2010.06512*, 2020.

P Balaresque and TE King. Human phenotypic diversity: An evolutionary perspective. *Current topics in developmental biology*, 119:349–390, 2016.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, pp. 3, 2016.

Fabiola Becerra-Riera, Annette Morales-González, and Heydi Méndez-Vázquez. A survey on facial soft biometrics for video surveillance and forensic applications. *Artificial Intelligence Review*, 52 (2):1155–1187, 2019.

Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 289–298, 2019.

Mary E Campbell and Lisa Troyer. The implications of racial misclassification by observers. *American Sociological Review*, 72(5):750–765, 2007.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Pierluigi Carcagnì, Marco Del Coco, Dario Cazzato, Marco Leo, and Cosimo Distante. A study on different experimental configurations for age, race, and gender estimation problems. *EURASIP Journal on Image and Video Processing*, 2015(1):1–22, 2015.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Stevie Chancellor, Eric PS Baumer, and Munmun De Choudhury. Who is the" human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32, 2019.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14980–14991, 2021.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554*, 2021.

Thomas Deselaers and Vittorio Ferrari. Visual and semantic similarity in imagenet. In *CVPR 2011*, pp. 1777–1784. IEEE, 2011.

Diana C Dima, Martin N Hebart, and Leyla Isik. A data-driven investigation of human action representations. *bioRxiv*, 2022.

Cynthia Feliciano. Shades of race: How phenotype and observer characteristics shape racial classification. *American Behavioral Scientist*, 60(4):390–419, 2016.

Jonathan B Freeman, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the part: Social status cues shape race perception. *PloS one*, 6(9):e25107, 2011.

Nelson Goodman. Seven strictures on similarity. 1972.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Mannat Singh, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision. *arXiv preprint arXiv:2202.08360*, 2022.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 501–512, 2020.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature human behaviour*, 4(11):1173–1185, 2020.

Martin N Hebart, Oliver Contier, Lina Teichmann, Adam Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker. Things-data: A multimodal collection of large-scale datasets for investigating object representations in brain and behavior. *bioRxiv*, 2022.

Mark E Hill. Race of the interviewer and perception of skin color: Evidence from the multi-city study of urban inequality. *American Sociological Review*, pp. 99–108, 2002.

Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

Emilie L Josephs, Martin N Hebart, and Talia Konkle. Emergent dimensions underlying human perception of the reachable world. *Journal of Vision*, 21(9):2154–2154, 2021.

Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1548–1558, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828, 2015.

Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

Zaid Khan and Yun Fu. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 587–597, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Arvindkumar Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference (BMVC)*, volume 1, pp. 3, 2021.

Anjana Lakshmi, Bernd Wittenbrink, Joshua Correll, and Debbie S Ma. The india face set: International and cultural boundaries impact face impressions and perceptions of category membership. *Frontiers in psychology*, 12:161, 2021.

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, pp. e1452, 2022.

Tom Leinster and Christina A Cobbold. Measuring diversity: the importance of species similarity. *Ecology*, 93(3):477–489, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods*, 47(4):1122–1135, 2015.

Debbie S Ma, Justin Kantner, and Bernd Wittenbrink. Chicago face database: Multiracial expansion. *Behavior Research Methods*, 53(3):1289–1300, 2021.

Nicolas Malevé. On the data set's ruins. *AI & SOCIETY*, pp. 1–15, 2020.

Arthur B Markman. Structural alignment in similarity and difference judgments. *Psychonomic Bulletin & Review*, 3(2):227–230, 1996.

Arthur B Markman and Dedre Gentner. Nonintentional similarity processing. *The new unconscious*, pp. 107–137, 2005.

John McCauley, Sobhan Soleymani, Brady Williams, John Dando, Nasser Nasrabadi, and Jeremy Dawson. Identical twins as a facial similarity benchmark for human facial recognition. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–5. IEEE, 2021.

Douglas L Medin, Robert L Goldstone, and Dedre Gentner. Respects for similarity. *Psychological review*, 100(2):254, 1993.

Thomas Nagel. *The view from nowhere*. oxford university press, 1989.

Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *European Conference on Computer Vision*, pp. 739–755. Springer, 2020.

Joshua C Peterson, Joshua T Abbott, and Thomas L Griffiths. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive science*, 42 (8):2648–2669, 2018.

Brett D Roads and Bradley C Love. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3547–3557, 2021.

Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pp. 0–1, 2020.

Eleanor Rosch. Cognitive representations of semantic categories. *Journal of experimental psychology: General*, 104(3):192, 1975.

Eleanor H Rosch. Natural categories. *Cognitive psychology*, 4(3):328–350, 1973.

Wendy D Roth. The multiple dimensions of race. *Ethnic and Racial Studies*, 39(8):1310–1338, 2016.

Myron Rothbart and Marjorie Taylor. Category labels and social reality: Do we view social categories as natural kinds? 1992.

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. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Amir Sadovnik, Wassim Gharbi, Thanh Vu, and Andrew Gallagher. Finding your lookalike: Measuring face similarity rather than face identity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2345–2353, 2018.

Craig A Sanders and Robert M Nosofsky. Training deep networks to construct a psychological feature space for a natural-object category domain. *Computational Brain & Behavior*, 3(3):229–251, 2020.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37, 2021.

Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Rebecca Pantofaru. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6: 2378023120967171, 2020.

Marshall H Segall, Donald Thomas Campbell, and Melville Jean Herskovits. *The influence of culture on visual perception*. Bobbs-Merrill Indianapolis, 1966.

Rosyl S Somai and Peter JB Hancock. Exploring perceived face similarity and its relation to image-based spaces: an effect of familiarity. *Journal of Vision*, 21(9):2149–2149, 2021.

Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 701–713, 2021.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 830–838, 2017.

Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5683–5692, 2019.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5265–5274, 2018.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 692–702, 2019.

Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*, pp. 25313–25330. PMLR, 2022.

Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision (ICCV)*, 2021.

Charles Y Zheng, Francisco Pereira, Chris I Baker, and Martin N Hebart. Revealing interpretable object representations from human behavior. In *International Conference on Learning Representations*, 2019.