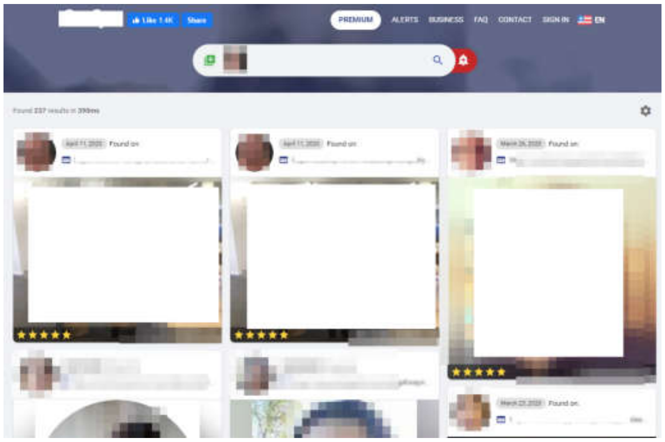
Supplementary material
Large image datasets: A pyrrhic win for computer vision?

Abeba Birhane*
School of Computer Science
Lero & University College Dublin, Ireland
abeba.birhane@ucdconnect.ie

Vinay Uday Prabhu*
UnifyID AI Labs
Redwood City, USA
vinay@unify.id

1. Risk of privacy loss via reverse search engines

As covered in the main paper, reverse image search engines[1] that facilitate face search such as [1] have gotten remarkably and worryingly efficient in the past year. For a small fee, anyone can use their portal or their API to run an automated process to uncover the *real-world* identities of the *humans of ImageNet* dataset. While both the men and women of imagenet dataset are under this risk, there is asymmetric risk here as the high NSFW classes such as `bra, bikini and maillot` are often the ones with higher female-to-men ratio (See Figure 5). Figure 1 showcases the snapshot image of one such reverse image search portal to demonstrate how easy it is for someone even with zero programming skills to just use their GUI to uncover the real world identities of the persons which can lead to catastrophic downstream risks such as blackmailing and other forms on online abuse as detailed in [34].



Found 237 results in 390ms

Figure 1: Snapshot of a popular reverse image search website

2. Quantitative auditing

In this section, we cover the details of performing the quantitative analysis on the ImageNet dataset including the following metrics: Person CAG (Count -Age - Gender) , NSFW scoring of the images, Semanticity and classification accuracy. The pre-trained models used in this endeavor are covered in Table 1. All of these analyses and the generated meta-datasets have been duly open sourced at `https://rb.gy/zccdps`. Figure 2 covers the details of all the jupyter notebooks authored to generate the datasets covered in Table 4.

2.1. Count, Age and Gender

In order to perform a human-centric census covering metrics such as count, age, and gender, we used the `InsightFace` toolkit for face analysis [19], that provided implementations of: `ArcFace` for deep face recognition [11] and `Retina-Face` for face localisation (bounding-box generation) [11]. We then combined the results of these models with the results obtained from [12] that used the `DEX` [39] model.

The results are as shown in Table 2 that captures the summary statistics for the `ILSVRC2012` dataset. In this table, the lower case n denotes the number of *images with persons* identified in them whereas the N indicates the *number of persons*[2]. The superscript indicates the algorithm used (`DEX` or `InsightFace` (if)) whereas the subscript has two fields: The train or validation subset indicator and the census gender-category. For example, $n^{(if)}_{val-O} = 3096$ implies that there were 3096 images in the ImageNet validation set (out of 50000) where the `InsightFace` models were able to detect a person's face.

As seen, the `InsightFace` model identified 101,070 persons across 83,436 images (including the train and validation subsets) which puts the prevalence rate of persons whose presence in the dataset exists sans explicit consent to be

[*] Equal contributions

[1] For example, PimEyes: `https://bit.ly/3bSKcZQ`

[2] The difference is simply on account of more than one person being identified by the model in a given image

Metric	Models used
Count, Age and Gender	`DEX` ([39]), `InsightFace` ([19]), RetinaFace [11], ArcFace [10]
NSFW-scoring	NSFW-MobileNet-V2-224 [17]
Semanticity	Glove [33], UMAP [29]
Classification Accuracy	Resent-50 [20], NasNet-mobile [47]

Table 1: Metrics considered and pre-trained models used

$N^{(dex)}_{train-O}$	$n^{(if)}_{train-O}$	$n^{(if)}_{val-O}$	$N^{(if)}_{train-O}$	$N^{(if)}_{val-O}$	$N^{(if)}_{train-W}$	$N^{(if)}_{train-M}$	$N^{(if)}_{val-W}$	$N^{(if)}_{val-M}$
132201	80340	3096	97678	3392	26195	71439	645	2307

Table 2: *Humans of the imagenet dataset*: How many?
Key: $\{n/N\}^{(\{dex/if\})}_{\{train/val\}-\{O/W/M\}}$.(O:Overall,W:Women,M: Men)

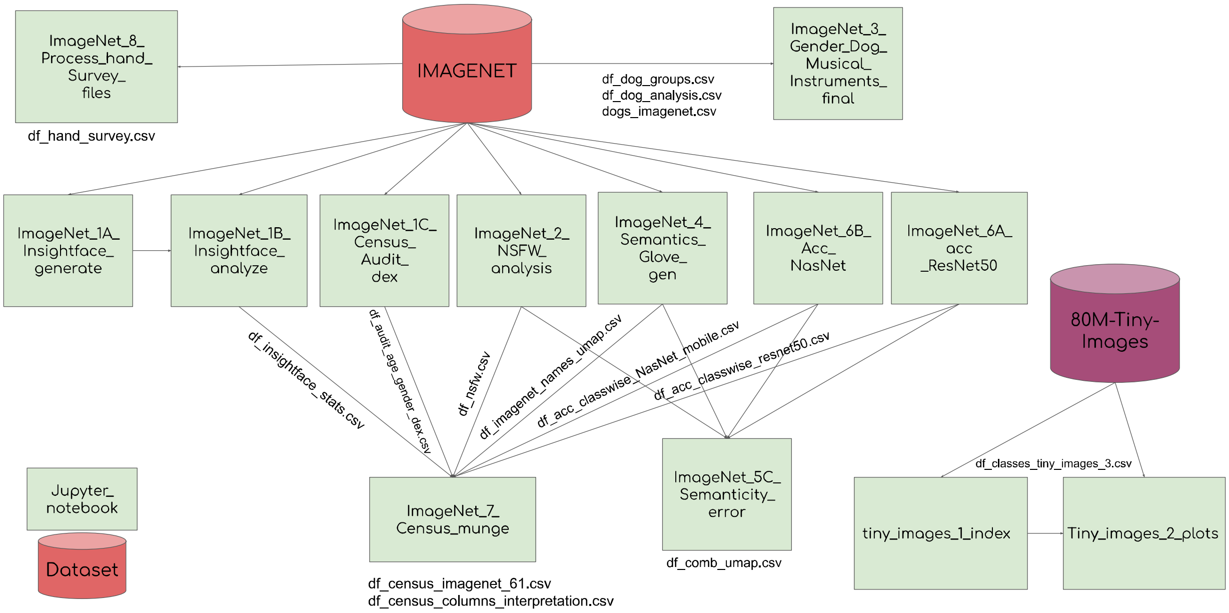

ImageNet_8_ Process_hand_ Survey_ files
df_hand_survey.csv
IMAGENET
df_dog_groups.csv
df_dog_analysis.csv
dogs_imagenet.csv
ImageNet_3_ Gender_Dog_ Musical_ Instruments_ final
ImageNet_1A_ Insightface_ generate
ImageNet_1B_ Insightface_ analyze
ImageNet_1C_ Census_ Audit_ dex
ImageNet_2_ NSFW_ analysis
ImageNet_4_ Semantics_ Glove_ gen
ImageNet_6B_ Acc_ NasNet
ImageNet_6A_ acc _ResNet50
80M-Tiny-Images
df_insightface_stats.csv
df_audit_age_gender_dex.csv
df_nsfw.csv
df_imagenet_names_umap.csv
df_acc_classwise_NasNet_mobile.csv
df_acc_classwise_resnet50.csv
df_classes_tiny_images_3.csv
Jupyter_ notebook
Dataset
ImageNet_7_ Census_munge
df_census_imagenet_61.csv
df_census_columns_interpretation.csv
ImageNet_5C_ Semanticity_ error
df_comb_umap.csv
tiny_images_1_index
Tiny_images_2_plots

Figure 2: Visualization of all the notebooks and dataset assets curated during the quantitative analysis

around 7.6% which is less aggressive compared to the 10.3% predicted by the DEX model, which has a higher identification false positive rate. An example of this can be seen in Fig 3 which showcases an example image with the bounding boxes of the detected persons in the image. Much akin to [12], we found a strong bias towards (older) male presence (73,746 with a mean age of 33.24 compared to 26,840 with a mean age of 25.58). At this juncture, we'd like to reemphasize that these *high accuracy* pre-trained models can indeed be highly error prone conditioned on the ethnicity of the person, as analyzed in [5, 12] and we'd like to invite the community to re-audit these images with better and more ethical tools (See Fig 4 for example of errors we could spot during the inference stage). Figure 6(a), presents the class-wise estimates of the number of persons in the dataset using the DEX and the InsightFace models. In Figure 6(b), we capture the variation in the estimates of count, gender and age of the `DEX` and the `InsightFace` models.

Before delving in to the discussions of the results obtained, we define the parameters that were measured. To begin, we denote ϕ_i to be the binary *face-present* indicator variable($\phi_i = \begin{cases} 1 & \text{if face present} \\ 0 & \text{otherwise} \end{cases}$) with regards to the image indexed i, (A) (in the superscripts) to be the algorithm used ($A \in \{\texttt{DEX}, \texttt{INSIGHTFACE}\}$), and N_c to be the number of images in the class c. Now, we define the class-level mean person count ($\eta_c^{(A)}$), mean-gender-skewness score ($\xi_c^{(A)}$) and mean-age ($\alpha_c^{(A)}$) to be,

class_number	label	mean_gender_audit	mean_age_audit	mean_nsfw_train
445	bikini, two-piece	0.18	24.89	0.859
638	maillot	0.18	25.91	0.802
639	maillot, tank suit	0.18	26.67	0.769
655	miniskirt, mini	0.19	29.95	0.62
459	brassiere, bra, bandeau	0.16	25.03	0.61

Table 3: Table of the 5 classes for further investigation that emerged from the NSFW analysis

file_name	shape	file_contents
df_insightface_stats.csv	(1000, 30)	24 classwise statistical parameters obtained by running the `InsightFace` model ([19]) on the ImageNet dataset
df_audit_age_gender_dex.csv	(1000, 12)	11 classwise (ordered by the wordnet-id) statistical parameters obtained from the json files (of the DEX paper) [39]
df_nsfw.csv	(1000, 5)	The mean and std of the NSFW scores of the train and val images arranged per-class. (Unnamed: 0: WordNetID of the class)
df_acc_classwise_resnet50.csv	(1000, 7)	Classwise accuracy metrics (& the image level preds) obtained by running the ResNet50 model on ImageNet train and Val sets
df_acc_classwise_NasNet_mobile.csv	(1000, 7)	Classwise accuracy metrics (& the image level preds) obtained by running the NasNet model on ImageNet train and Val sets
df_imagenet_names_umap.csv	(1000, 5)	DF with 2D UMAP embeddings of the Glove vectors of the classes of the ImageNet dataset
df_census_imagenet_61.csv	(1000, 61)	The MAIN census dataframe covering class-wise metrics across 61 parameters, all of which are explained in df_census_columns_interpretation.csv
df_census_columns_interpretation.csv	(61, 2)	The interpretations of the 61 metrics of the census dataframe above!
df_hand_survey.csv	(61, 3)	Dataframe contaimning the details of the 61 images unearthed via hand survey (Do not pay heed to 61. it is a mere coincidence)
df_classes_tiny_images_3.csv	(75846, 3)	Dataframe containing the class_ind, class_name (wordnet noun) and n_images
df_dog_analysis.csv	(7, 4)	Dataframe containing breed, gender_ratio and survey result from the paper Breed differences in canine aggression'

Table 4: Meta datasets curated during the audit processes

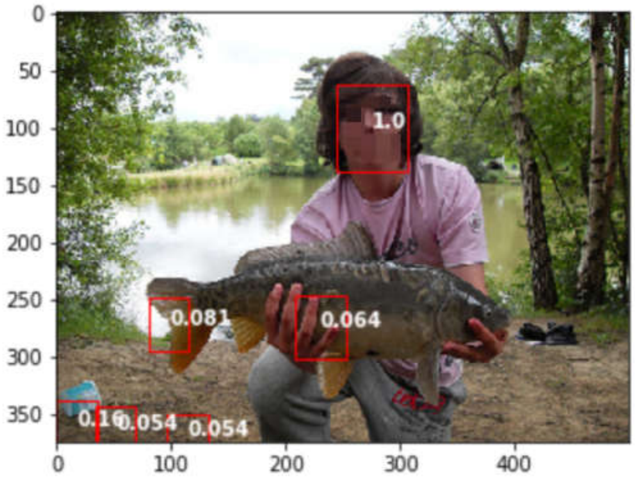


Figure 3: An example image with the output bounding boxes and the confidence scores of the humans detected in the image by the `DEX` model([39])

$$\eta_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i], \alpha_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i] a_i^{(A)} \ and$$

$$\xi_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i]\left(\frac{g_i^{(A)} - \mu_c^{(A)}}{\sigma_c^{(A)}}\right)^3.$$

Here, $\mu_c^{(A)}$ and $\sigma_c^{(A)}$ represent the mean and standard-deviation of the gender-estimate of the images belonging to class c and estimated by algorithm (A) respectively.

With regards to the first scatter-plot in Figure 6(b), we observe that the estimated class-wise counts of persons ($\eta_c^{(A)}$) detected by the `DEX` and `InsightFace` models in the images were in strong agreement ($Pearson - r = 0.973(0.0)$) which helps to further establish the global person prevalence rate in the images to be in the order of $7.6 - 10.3\%$. These scatter-plots constitute Figure 12 of the dataset audit card

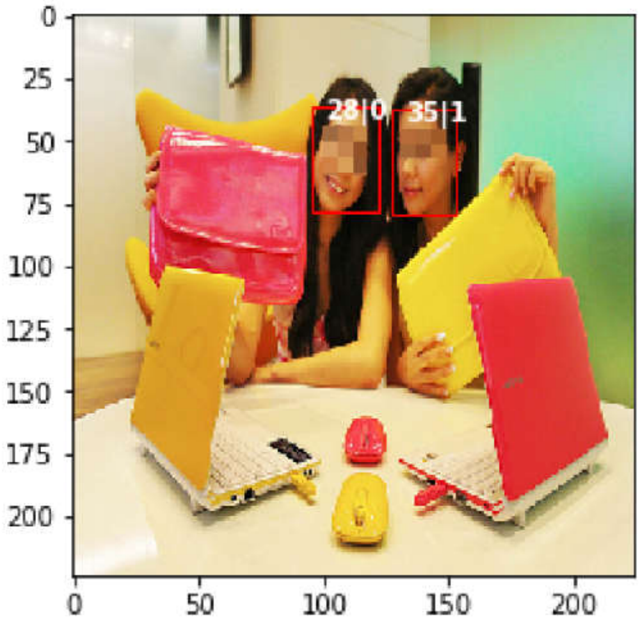


Figure 4: An example image with the output bounding boxes and the estimated ages/ (binarized) genders of the persons detected in the image by the `InsightFace` model. (Here 0: female and 1: Male)

(Figure 15).
Now, we would like to draw the attention of the reader towards the weaker correlation ($Pearson - r = 0.723(0.0)$) when it came to gender-skewness ($\xi_c^{(A)}$) and the mean age-estimates ($\alpha_c^{(A)}$; $Pearson - r = 0.567(0.0)$) scatter-plots in Figure 6(b). Given that the algorithms used are state-of-the-art with regards to the datasets they have been trained on (see [39] and [19]), the high disagreement on a neutral dataset like ImageNet exposes the frailties of these algorithmic pipelines upon experiencing population shifts in the test dataset. This, we believe lends further credence to the studies that have demonstrated poor reliability of these so-termed accurate models upon change of the underlying demographics (see [12] and [5]) and also supports the movement to move away from gender classification of account of not just the moral and ethical repugnance of the inherent task itself but also on account of the *scientific validity* argument as well [43].

2.2. NSFW scoring aided misogynistic imagery hand-labeling

Previous journalistic efforts (see [36]) had revealed the anecdotal presence of strongly misogynistic content in the imagenet dataset, specifically in the categories of `beach-voyeur-photography, upskirt images, verifiably pornographic and exposed private-parts`. These specific four categories have been well researched in digital criminology and intersectional feminism (see [22, 28, 34, 35]) and have formed the backbone of several legislations all over the world (see [27],[18]). In order to help generate a hand labelled dataset of these images amongst more than 1.3 million images, we used a hybrid human-in-the-loop approach where we first formed a smaller subset of images from image classes filtered using a model-annotated NSFW-average score as a proxy. For this, we used the `NSFW-Mobilenet-v2` model [17] which is an image-classification model with the output classes being `[drawings, hentai, neutral, porn, sexy]`. We defined the NSFW score of an image by summing up the softmax values of the `[hentai, porn, sexy]` subset of classes and estimated the mean-NSFW score of all of the images of a class to obtain the results portrayed in Figure 7. In Figure 7(a), we see the scatter-plot of the mean-NSFW scores plotted against the mean-gender scores (obtained from the DEX model estimates) for the 1000 imagenet classes. We then found five natural clusters upon using the *Affinity Propagation* algorithm [16]. Given the `0:FEMALE|1:MALE` gender assignments in the model we used (see [12]), classes with lower mean-gender scores allude towards a *women-majority class*). The specific details of the highlighted cluster in the scatter-plot in Figure 12(a) are displayed in Table 3. Further introducing the age dimension (by way of utilising the mean-age metric for each class), we see in Figure 12(b), that the classes with the highest NSFW scores were those where the dominating demographic was that of young women. With this shortlisting methodology, we were left with approximately 7000 images which were then hand labelled by a team of five volunteers (three male, two female, all aged between 23-45) to curate a list of 61 images where there was complete agreement over the 4 class assignment. We have duly open-sourced the hand-curated list (see Table 5), and the summary results are as showcased in Figure 8. In sub-figure Figure 8(a), we see the cross-tabulated class-wise counts of the four categories of images[3] across the imagenet classes and in Figure 8(b), we present the histogram-plots of these 61 hand-labelled images across the imagenet classes. As seen, the `bikini, two-piece` class with a mean NSFW score of 0.859 was the *main* image class with 24 confirmed `beach-voyeur` pictures.
Here, we would like to strongly reemphasise that we are disseminating this list as a community resource so as to facilitate further scholarly debate and also, if need be, allow scholars in countries where incriminating laws (See [13]) may exist, to deal with in the correct topical way they deem fit. We certainly admit to the primacy of **context** in which the *objectionable* content appears. For example, the image `n03617480_6206.jpeg` in the class `n03617480 - kimono` that contained genital exposure, turned out to be a photographic bondage art piece shot by Nobuyoshi Araki[31] that straddles the fine line between *scopophilic eroticism* and pornography. But, as explored in [13], the mere possession of a digital copy of this picture would be punishable by law in many nation states and we believe that these factors have to be considered while disseminating a large scale image dataset or should be detailed as caveats in the dissemination document.

2.2.1 NSFW and semanticity of classes

We also analyzed the relationship between the semanticity of classes and NSFW scores obtained. Firstly, we obtained a representative word for each of the 1000 class labels in ILSVRC2012 and used [33] to generate dense word-vector *Glove* embeddings in 300-D. Further, in order to generate the 2D/3D scatter-plots in Figure 5, we used the UMAP [29] algorithm to perform dimensionality reduction. `df_imagenet_names_umap.csv` contains the 2D UMAP embeddings of the resultant Glove vectors of the classes that are then visualized in Figure 5 (a). In Figure 5 (b), we see the 3D surface plot of the 2D UMAP semantic dimensions versus the NSFW scores. As seen, it is peaky in specific points of the semantic space of the label categories mapping to classes such as `brassier, bikini and`

[3] This constitutes Figure 13(in the data audit card)

maillot.

2.3. Dogs to musical instruments: Co-occurrence based gender biases

Social and intersectional biases prevalent in the society, do seep into datasets and the statistical models trained on them. In the context of Natural Language Processing (NLP), the framework of lexical co-occurrence has been harnessed to tease out these biases, especially in the context of gender biases. In [42], the authors analyzed stereotypically male occupation words (that they termed as *M-biased* words) as well as stereotypically female occupation words (*F-biased* words) in large text corpora and the ensuing downstream effects when used to generate contextual word representations in SoTA models such as such as BERT and GPT-2. Further, in [38], direct normalized co-occurrence associations between the word and the representative concept words were proposed as a novel corpus bias measurement method, and it's efficacy was demonstrated with regards to the actual gender bias statistics of the U.S. job market and it's estimates measured via the text corpora. In the context of the imagenet dataset, we investigated if such co-occurrence biases do exist in the context of human co-occurrence in the images. Previously, in [41], the authors had explored the biased representation learning of an imagenet trained model by considering the class basketball where images containing black persons were deemed *prototypical*. Here, we tried to investigate if the gender of the person co-occurring in the background alongside the non-person class was skewed along the lines that it is purported to be in related academic work. We performed these investigations in the context of person-occurrence with regards to dog-breeds as well as musical instruments. Presented in Figure 9 is the gender-conditioned violin plot of co-occurrence of the person in the image alongside the group of dog-breed in the imagenet dataset. We obtained these measurement in two phases. In the first phase, we grouped the 120 imagenet classes of dog-breeds in to the following 7 groups: [Toy, Hound ,Sporting, Terrier, Non-Sporting, Working, Herding] following the formal American Kennel Club[4] (AKC) groupings (see [7]). The remaining breeds not in the AKC list were placed into Unknown. Once grouped, we computed the gender-conditioned population spreads of person-concurrence using the mean-gender value of the constituent image classes obtained estimated from [12]. Prior literature, both academic (See [25],[37]) and non-academic [26, 30] have studied the nexus between the perceived *manliness* of dog groups and the ownership gender. These stereotypical associations were indeed reflected in the person co-occurrence gender distributions in Figure 9(a), where we see that the so perceived *masculine* dog groups belonging to the set [Non-Sporting, Working, Herding] had a stronger male-gender co-occurrence bias. In similar vein, in Figure 9(b) we present the gender-skewness $(\xi_c^{(DEX)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i]\left(\frac{g_i^{(DEX)}-\mu_c}{\sigma_c}\right)^3)$ variation amongst the co-occurring persons across the 17 imagenet musical instrument classes. Works such as [8], [46] and [4] have explored in depth, the gender biases there exist in musical instrument selection. As stated in [44], instruments such as the cello, oboe, flute and violin have been stereotypically tagged to be *feminine* whereas instruments such as drum, banjo, trombone, trumpet and the saxophone were the so-termed *masculine* instruments in the western context. While these stereotypes are valid in a spatio-temporally constrained context, the *west-centrism-bias* [5] of the search engine used to curate the dataset has resulted in the mirroring of these topical real-world association biases. As seen in Figure 9(b), harp, cello,oboe, flute and violin indeed had the strongest *pro-women* bias where as drum, banjo, trombone, trumpet and saxophone were the classes with the strongest *male leaning* skewness scores.

2.4. Classes containing pictures of infants

We found this category to be particularly pertinent both under the wake of strong legislation's protecting privacy of children's' digital images as well as the extent of it. We found pictures of infants and children across the 30 image classes (and possibly more): ['bassinet', 'cradle', 'crib', 'bib', 'diaper', 'bubble', 'sunscreen', 'plastic bag', 'hamper', 'seat belt', 'bath towel', 'mask', 'bow-tie', 'tub', 'bucket', 'umbrella', 'punching bag', 'maillot - tank suit', 'swing', 'pajama', 'horizontal bar', 'computer keyboard', 'shoe-shop', 'soccer ball', 'croquet ball', 'sunglasses', 'ladles', 'tricycle - trike - velocipede', 'screwdriver', 'carousel']. What was disappointing was the prevalence of entire classes such as 'bassinet', 'cradle', 'crib' and 'bib' that had very high density of images of infants. We believe this might have legal ramifications as well. For example, Article 8 of the European Union General Data Protection Regulation (GDPR), specifically deals with the *conditions applicable to child's consent in relation to information society services* [32] The associated *Recital 38* states

[4] AKC claims that registered breeds are *assigned to one of seven groups representing characteristics and functions the breeds were originally bred for.*

[5] See https://www.kaggle.com/c/inclusive-images-challenge

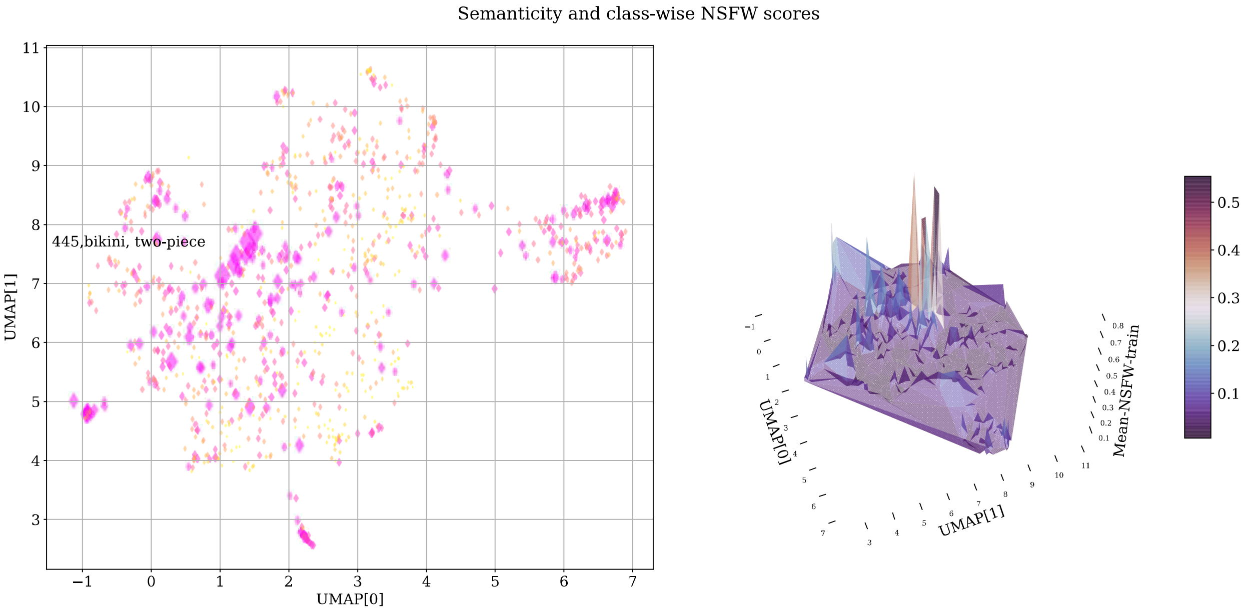


Figure 5: Figure showcasing the relationship between the semanticity of classes and the class-wise mean NSFW scores

verbatim that *Children merit specific protection with regard to their personal data, as they may be less aware of the risks, consequences and safeguards concerned and their rights in relation to the processing of personal data. Such specific protection should, in particular, apply to the use of personal data of children for the purposes of marketing or creating personality or user profiles and the collection of personal data with regard to children when using services offered directly to a child.* Further, **Article 14** of GDPR explicitly states: *Information to be provided where personal data have not been obtained from the data subject.* We advocate allying with the legal community in this regard to address the concerns raised above.

2.5. Blood diamond effect in models trained on this dataset

Akin to the *ivory carving-illegal poaching* and *diamond jewelry art-blood diamond* nexuses, we posit there is a similar moral conundrum at play here and would like to instigate a conversation amongst the neural artists in the community. The emergence of tools such as BigGAN [**?**] and GANbreeder [40] has ushered in an exciting new flavor of generative digital art [3], generated using deep neural networks (See [24] for a survey). A cursory search on twitter[6] reveals hundreds of interesting art-works created using BigGANs. There are many detailed blog-posts[7] on generating neural art by beginning with seed images and performing nifty experiments in the latent space of BigGANs. At the point of authoring this paper, (6/26/2020, 10:34 PM PST), users on the *ArtBreeder* app[8] had *generated* 64683549 images. Further, *Christie's*, the British auction house behemoth, recently hailed the selling of the neural network generated *Portrait of Edmond Belamy* for an incredible $432,500 as *signalling the arrival of AI art on the world auction stage*[6]. Given the rapid growth of this field, we believe this to be the right time to have a conversation about a particularly dark ethical consequence of using such frameworks that entail models trained on the *ImageNet* dataset which has many images that are pornographic, non-consensual, voyeuristic and also entail underage nudity. We argue that this lack of consent in the seed images used to train the models trickles down to the final art-form in a way similar to the blood-diamond syndrome in jewelry art [14].

An example: Consider the neural art image in Fig 10 we generated using the *GanBreeder* app. On first appearance, it is not very evident as to what the constituent *seed* classes are that went into the creation of this neural artwork image. When we solicited volunteers online to critique the artwork (See the collection of responses in Table 6), none had an inkling regarding a rather sinister trickle down effect at play here. As it turns out, we craftily generated this image using hand-picked specific instances of *children* images emanating from what we will showcase are two problematic *seed* image classes: Bikini and Brassiere. More specifically, for this particular image, we set the *Gene weights* to be: [*Bikini*: 42.35, *Brassiere*: 31.66, *Comic Book* - 84.84]. We'd like to strongly emphasize at this juncture that the problem does

[6] https://twitter.com/hashtag/biggan?lang=en

[7] https://rb.gy/pr9pwb

[8] https://ganbreeder.app

not emanate from a visual patriarchal mindset [2], whereby we associate female undergarment imagery to be somehow unethical, but the root cause lies in the fact that many of the images were curated into the dataset (at least with regards to the 2 above mentioned classes) were voyeuristic, pornographic, non-consensual and also entailed underage nudity.

2.6. Error analysis

Given how besotted the computer vision community is with regards to classification accuracy metrics, we decided to understand what happens to the class-wise top-5 accuracies in those classes where humans co-occur asymmetrically between the training and validation sets. For this, we performed inference using the ResNet50 [20] and NasNet [47] models and sorted all the 1000 classes as per the $N_{train}^{persons}/N_{val}^{persons}$ ratios (termed *human-delta* in the figure) and compared their accuracies with regards to the general population (amongst the 1000 classes). As gathered from Figure 11, we saw a statistically significant drop in top-5 accuracies ($T - test \in (-3.87, -3.06)$) for the *top-25 human-delta classes*, thereby motivating that even for the purveyors of *scientism* fuelled pragmatism, there is motivation here to pay heed to the problem of humans in images. We'd like to reemphasize that we are most certainly not advocating this to be the *prima causa* for instigating a cultural change in the computer vision community, but are sharing these resources and nuances for further investigation.

3. Broader impact statement and a wish list

We authored this paper with an aspiration to strive for a broader impact in terms of instigating a fundamental change in the way institutions, both academic and industry, curate large scale image dataset. Through the course of this work, we solicited and incorporated feedback from scholars in the field whose pointed us towards three valid critiques that we'd like to address first. To begin with, we solemnly acknowledge the moral paradox in our use of pre-trained gender classification models for auditing the dataset and duly address this in the previous section. Secondly, as covered in Section on *threat landscape*, we also considered the risks of the possible *Streissand effect* with regards to deanonymization of the persons in the dataset that ultimately lead us to not dive further into the quantitative or qualitative aspects of our findings in this regard, besides conveying a specific example via email to the curator of the dataset from which the deanonymization arose. Thirdly, we are wary of the risk of this body of critique being misappropriated to further a misogynistic agenda by the regressive components of the machine learning community as we feel there is a constant under-celebration of the impact of dataset curation when juxtaposed with SoTA metric achieving architecture tweaking. This is often pitched as an *AlexNet versus ImageNet* styled strawman narrative[9] , one that needs to be tackled on war footing. In this regard, we would like to explicitly acknowledge the gargantuan effort that went into curation of these massive datasets and do not in the slightest, aspire to undermine the brilliance of the effort. With these caveats firmly in tow, we now proceed to conclude with the following Wish list of the impact we hope this work may bring about.

3.1. Proactive approach over reactive course corrections

We aspire to see the institutions curating these large scale datasets to be proactive in establishing the primacy of ethics in the dataset curation process and not just reacting to exposes and pursing posthoc course corrections as an afterthought. We'd be well served to remind ourselves hat it took the community 11 years to go from the first peer-reviewed dissemination [9] of the imagenet dataset to achieving the first meaningful course correction in [45] whereas the *number of floating-point operations required to train a classifier to AlexNet-level performance on ImageNet had decreased by a factor of 44x between 2012 and 2019* [23]. This, we believe, demonstrates where the priorities lie and this is precisely where we seek to see the most impact.

3.2. Bluewashing of AI ethics and revisiting the enterprise of Big data

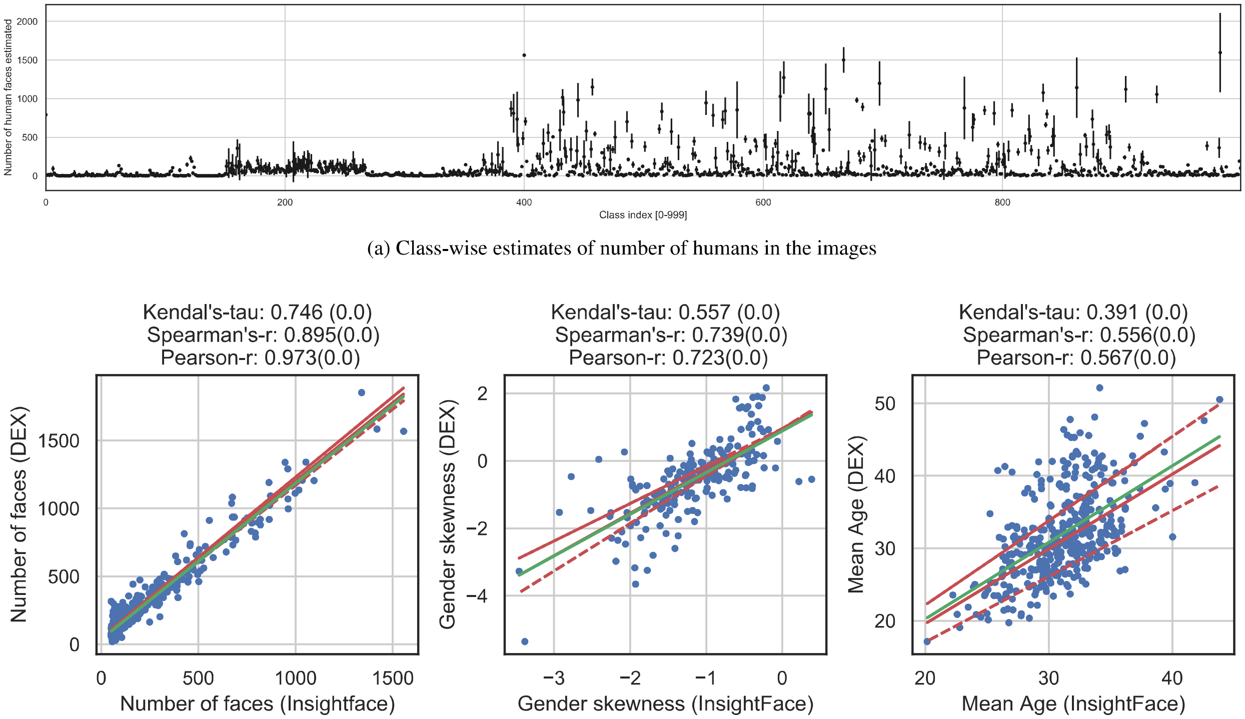
At the outset, we question if *Big Data* can ever operate in a manner that caters for marginalized communities - those disproportionately impact by algorithmic injustice. Automated large scale data harvesting forays, by their very volition, tend be *BIG*, in the sense that they are inherently prone to **B**ias, are **I**mperceptive to the lessons of human condition and recorded history of vulnerable people and **G**uileful to exploit the loopholes of legal frameworks that allow siphoning off of lived experiences of disfranchised individuals who have little to no agency and recourse to contend Big Data practices. Both collective silence and empty lip service [10] i.e. caricatured appropriations of ethical transgressions entailing *ethics shopping*, *ethics bluewashing*, *ethics lobbying*, *ethics dumping* and *ethics shirking* [15] cause harm and damage. Given that these datasets emerged from institutions such as Google, Stanford, NYU and MIT, all with substantial number of staff researching AI ethics and policy, we cannot help but feel that this hints towards not just compartmentalization and fetishization of ethics as a *hot topic* but also shrewd usage of the ethicists as agents of *activism outsourcing*.

3.3. Arresting the creative commons loot

As covered in the main paper, we could like to see this trend of using the creative commons loophole as an excuse

[9] https://bit.ly/3gXnQt0

[10] https://www.media.mit.edu/articles/beware-corporate-machinewashing-of-ai/



(a) Class-wise estimates of number of humans in the images

(b) Scatter-plots with correlations covering the cardinality, age and gender estimates

Figure 6: Juxtaposing the results from the `DEX` and the `InsightFace` models

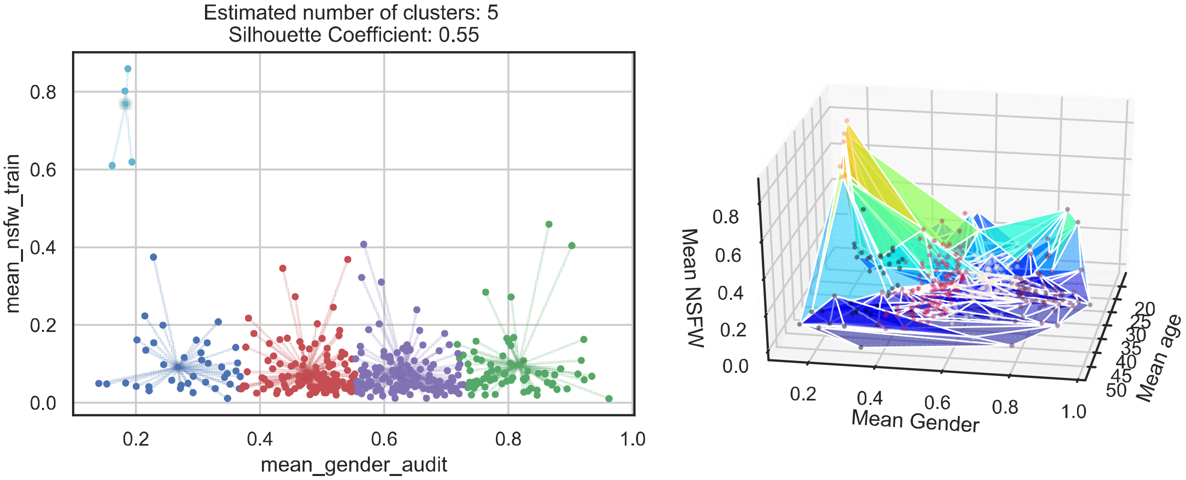
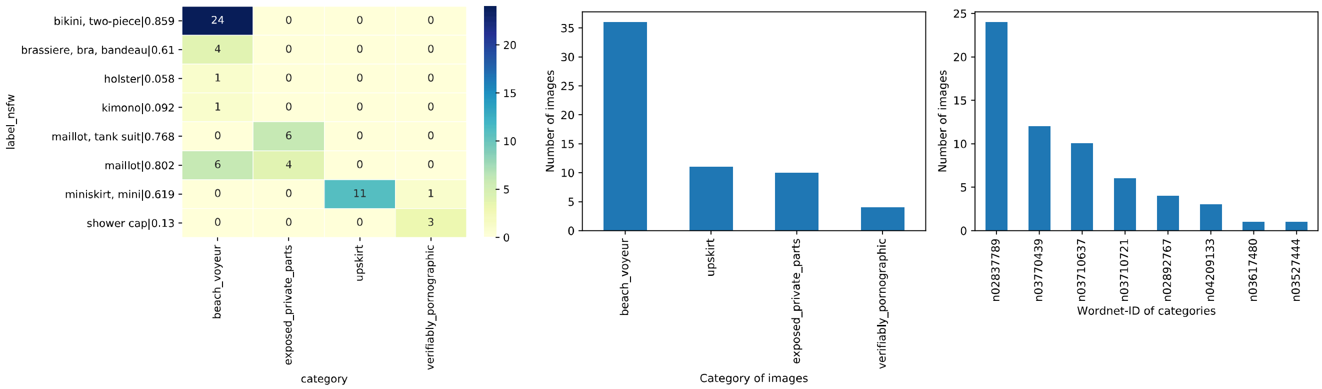


Figure 7: Class-wise cross-categorical scatter-plots across the age, gender and NSFW score estimates

for circumventing the difficult terrain of informed consent. We should as a field, aspire to treat consent in the same rigorous way that researchers and practitioners in fields such as anthropological studies or medical studies. In this work, we have sought to draw the attention of the Machine Learning community towards the societal and ethical implications of large scale datasets, such as the problem of non-consensual images and the oft-hidden problems of categorizing people. We were inspired by the adage of *Secrecy begets tyranny*[11] and wanted to issue this as a call to the Machine Learning community to pay close attention to the direct and indirect impact of our work on society, especially on vulnerable

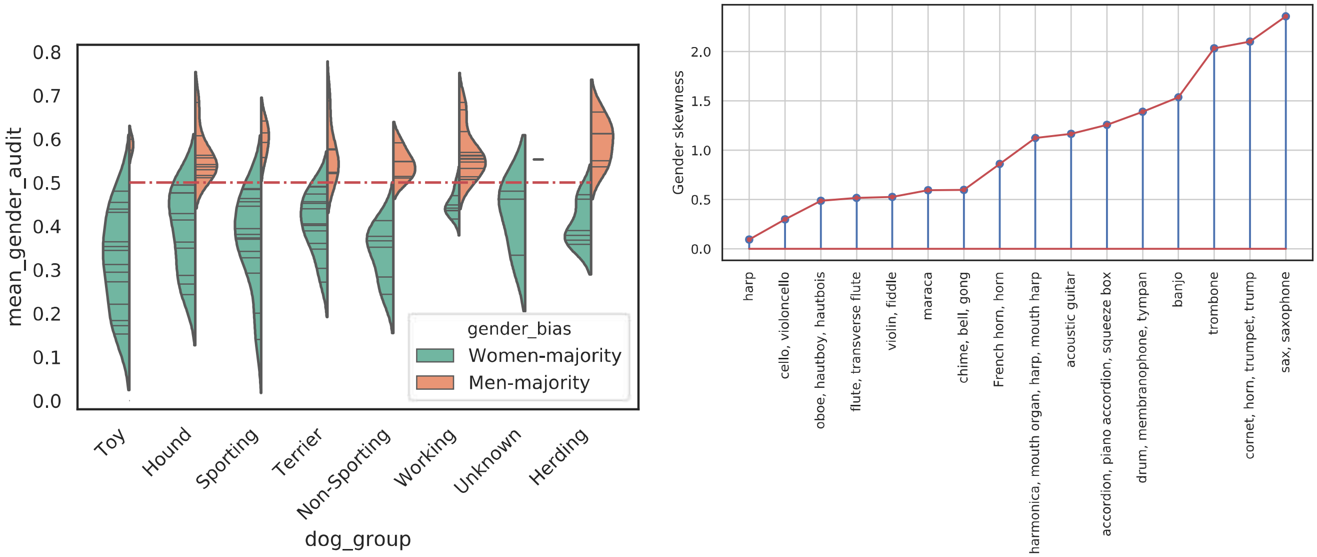
[11] From Robert A. Heinlein's 1961 science fiction novel titled *Stranger in a Strange Land [21]*



(a) Cross-tabulated grid-plot of the co-occurrence of the imagenet classes with the hand-labelled categories

(b) Histogram-plots of the hand-labelled images

Figure 8: Plots showcasing the statistics of the hand-survey across the `beach-voyeur, exposed-private-parts, upskirt, verifiably-pornographic` image categories



(a) Categorized violin plot demonstrating the class-wise mean gender scores across the dog-breed image groups

(b) Gender skewness scores across the different musical instrument image classes

Figure 9: Plots showcasing the *human co-occurrence* based gender-bias analysis

groups. We hope this work contributes to raising awareness and adds to a continued discussion of ethics in Machine Learning, along with many other scholars that have been elucidating algorithmic bias, injustice, and harm.

References

[1] Face search • pimeyes. `https://pimeyes.com/en/`, May 2020. (Accessed on 05/04/2020).

[2] Stephanie Baran. Visual patriarchy: Peta advertising and the commodification of sexualized bodies. In *Women and Nature?*, pages 43–56. Routledge, 2017.

[3] Margaret A Boden and Ernest A Edmonds. What is generative art? *Digital Creativity*, 20(1-2):21–46, 2009.

[4] Claudia Bullerjahn, Katharina Heller, and Jan Hoffmann. How masculine is a flute? a replication study on gender stereotypes and preferences for musical instruments among

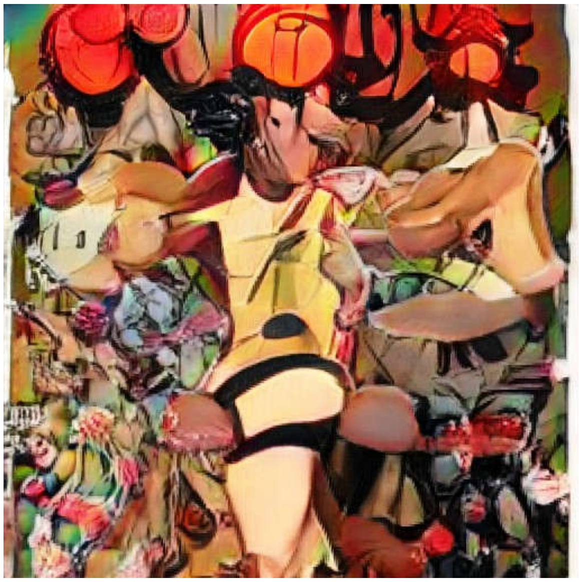


Figure 10: An example neural art image generated by the author using the *GanBreeder* app [Gene weights: Bikini: 42.35, Brassiere: 31.66, Comic Book - 84.84]

young children. In *Proceedings of the 14th International Conference on Music Perception and Cognition*, pages 5–9, 2016.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[6] Christies. Is artificial intelligence set to become art's next medium?, 2019. [Online; accessed 9-8-2019].

[7] American Kennel Club. List of breeds by group – american kennel club. `https://www.akc.org/public-education/resources/general-tips-information/dog-breeds-sorted-groups/`, Jan 2019. (Accessed on 05/31/2020).

[8] Judith K Delzell and David A Leppla. Gender association of musical instruments and preferences of fourth-grade students for selected instruments. *Journal of research in music education*, 40(2):93–103, 1992.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[11] Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arxiv*, 2019.

[12] Chris Dulhanty and Alexander Wong. Auditing imagenet: Towards a model-driven framework for annotating demographic attributes of large-scale image datasets. *arXiv preprint arXiv:1905.01347*, 2019.

[13] S. Durham. *Opposing Pornography: A look at the Anti-Pornography Movement*. Lulu.com, 2015.

[14] Julie L Fishman. Is diamond smuggling forever-the kimberley process certification scheme: The first step down the long road to solving the blood diamond trade problem. *U. Miami Bus. L. Rev.*, 13:217, 2004.

[15] Luciano Floridi. Translating principles into practices of digital ethics: five risks of being unethical. *Philosophy & Technology*, 32(2):185–193, 2019.

[16] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[17] Bedapudi Praneeth Gant Laborde. Nsfw detection machine learning model. `https://github.com/GantMan/nsfw_model`, Jan 2019. (Accessed on 05/31/2020).

[18] Alisdair A Gillespie. Tackling voyeurism: Is the voyeurism (offences) act 2019 a wasted opportunity? *The Modern Law Review*, 82(6):1107–1131, 2019.

[19] Jia Guo and Jiankang Deng. deepinsight/insightface: Face analysis project on mxnet. `https://github.com/deepinsight/insightface`, May 2020. (Accessed on 05/31/2020).

[20] 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*, pages 770–778, 2016.

[21] Robert A Heinlein. *Stranger in a strange land*. Hachette UK, 2014.

[22] Nicola Henry, Anastasia Powell, and Asher Flynn. Not just 'revenge pornography': Australians' experiences of image-based abuse. *A Summary Report, RMIT University, May*, 2017.

[23] Danny Hernandez and Tom B. Brown. Measuring the algorithmic efficiency of neural networks, 2020.

[24] Aaron Hertzmann. Aesthetics of neural network art. *arXiv preprint arXiv:1903.05696*, 2019.

[25] Elizabeth C Hirschman. Consumers and their animal companions. *Journal of consumer research*, 20(4):616–632, 1994.

[26] Jessica Macias. 10 manly dog breeds | petmd | petmd. `https://www.petmd.com/dog/slideshows/breeds/manly-dog-breeds?view_all=1`, Jun 2014. (Accessed on 06/01/2020).

[27] Clare McGlynn and Erika Rackley. More than revenge porn: image-based sexual abuse and the reform of irish law. *Irish probation journal.*, 14:38–51, 2017.

[28] Clare McGlynn, Erika Rackley, and Ruth Houghton. Beyond revenge porn: The continuum of image-based sexual abuse. *Feminist Legal Studies*, 25(1):25–46, 2017.

[29] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[30] Dom Naish. 20 most manly dog breeds for guys. `https://topdogtips.com/manly-dog-breeds/`, Aug 2017. (Accessed on 06/01/2020).

[31] Manamai Ozaki et al. Shashinjinsei: Nobuyoshi araki's photo journey art and not or pornography. *Art Monthly Australia*, (211):17, 2008.

[32] European Parliament and of the Council. Eur-lex - 32016r0679 - en - eur-lex. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`, Apr 2016. (Accessed on

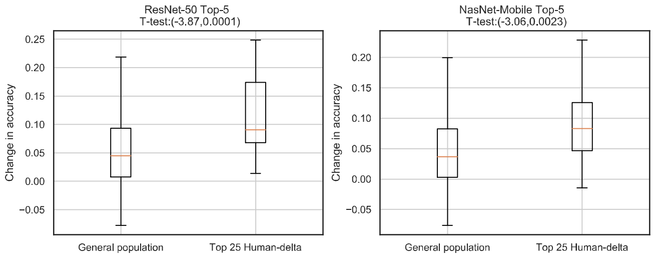


Figure 11: On accuracy variations and *human delta*

04/30/2020).

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[34] Anastasia Powell. Configuring consent: Emerging technologies, unauthorized sexual images and sexual assault. *Australian & New Zealand journal of criminology*, 43(1):76–90, 2010.

[35] Anastasia Powell, Nicola Henry, and Asher Flynn. Image-based sexual abuse. In *Routledge handbook of critical criminology*, pages 305–315. Routledge, 2018.

[36] Katyanna Quach. Inside the 1tb imagenet data set used to train the world's ai: Naked kids, drunken frat parties, porno stars, and more • the register. `https://www.theregister.co.uk/2019/10/23/ai_dataset_imagenet_consent/`, Oct 2019. (Accessed on 05/01/2020).

[37] Michael Ramirez. "my dog's just like me": Dog ownership as a gender display. *Symbolic Interaction*, 29(3):373–391, 2006.

[38] Navid Rekabsaz, James Henderson, Robert West, and Allan Hanbury. Measuring societal biases in text corpora via first-order co-occurrence. *arXiv:1812.10424 [cs, stat]*, Apr 2020. arXiv: 1812.10424.

[39] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[40] Joel Simon. Artbreeder. `https://www.artbreeder.com/about`, Jun 2020. (Accessed on 07/06/2020).

[41] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

[42] Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems*, pages 13209–13220, 2019.

[43] Sarah Myers West, Meredith Whittaker, and Kate Crawford. Discriminating systems. `https://ainowinstitute.org/discriminatingsystems.html`, 2019.

[44] Elizabeth R Wrape, Alexandra L Dittloff, and Jennifer L Callahan. Gender and musical instrument stereotypes in middle school children: Have trends changed? *Update: Applications of Research in Music Education*, 34(3):40–47, 2016.

[45] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 547–558, 2020.

[46] Jason Zervoudakes and Judith M Tanur. Gender and musical instruments: Winds of change? *Journal of Research in Music Education*, pages 58–67, 1994.

[47] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

Dataset audit card - ImageNet (Reproduced)

Census audit statistics

- 83436 images with 101070 − 132201 persons (Models: DEX ([39]), InsightFace ([19]))
- Mean-age (male): 33.24 (Female):25.58 (RetinaFace [11], ArcFace [10])
- Confirmed misogynistic images: 62. Number of classes with infants: 30
- ($\mu_c^{(A)}$ and $\sigma_c^{(A)}$: Mean and standard-deviation of the gender-estimate of images in class c estimated by algorithm (A).)

Metrics: Class-level mean count ($\eta_c^{(A)}$), mean gender skewness ($\xi_c^{(A)}$) and mean-age ($\alpha_c^{(A)}$):

$$\eta_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i], \alpha_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i] a_i^{(A)} \ and$$

$$\xi_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i]\left(\frac{g_i^{(A)} - \mu_c^{(A)}}{\sigma_c^{(A)}}\right)^3$$

$$\phi_i = \begin{cases} 1 & \text{if face present} \\ 0 & \text{otherwise} \end{cases} \text{ in } i^{th} \text{ image.}$$

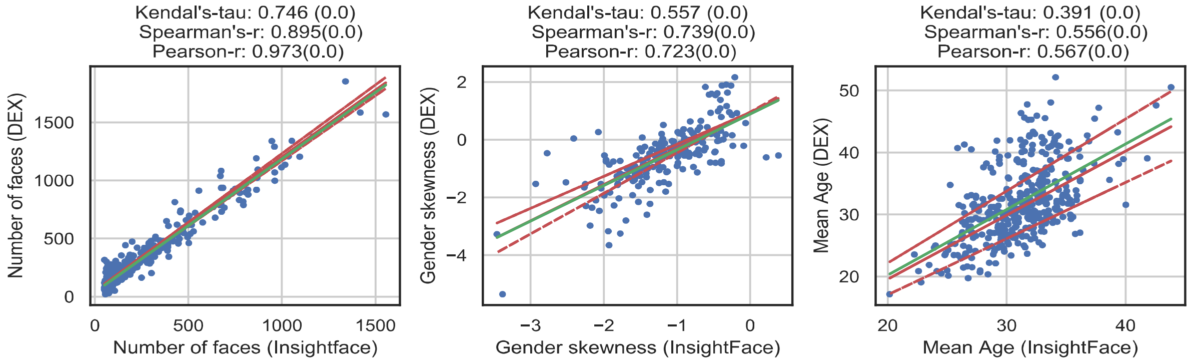


Figure 12: Class-wise cross-categorical scatter-plots across the cardinality, age and gender scores

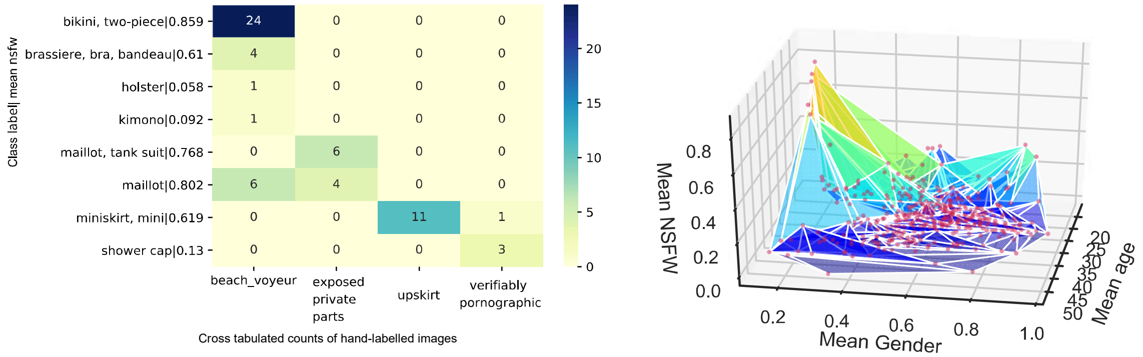


Figure 13: Statistics and locationing of the hand-labelled images

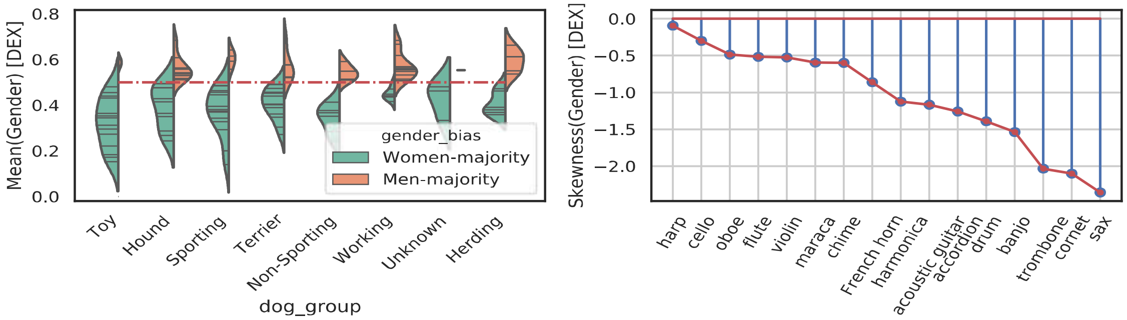


Figure 14: Known *human co-occurrence* based gender-bias analysis

Figure 15: Dataset audit card for the ImageNet dataset

wordnet_id	label	mean_nsfw_train	category	file_names
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_11383.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_12451.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_13794.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_14133.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_15158.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_15170.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_15864.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_17.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_17291.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_17410.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_18107.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_18124.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_18260.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_20096.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_22044.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_283.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_3414.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_3536.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_4.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_5713.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_9181.JPEG
n02837789	bikini, two-piece	0.859369	beach_voyeur	n02837789_9859.JPEG
n02837789	bikini, two-piece	0.859369	exposed_private_parts	n02837789_17069.JPEG
n02837789	bikini, two-piece	0.859369	exposed_private_parts	n02837789_19619.JPEG
n02892767	brassiere, bra, bandeau	0.610233	exposed_private_parts	n02892767_19629.JPEG
n02892767	brassiere, bra, bandeau	0.610233	exposed_private_parts	n02892767_3235.JPEG
n02892767	brassiere, bra, bandeau	0.610233	upskirt	n02892767_17717.JPEG
n02892767	brassiere, bra, bandeau	0.610233	verifiably_pornographic	n02892767_5914.JPEG
n03527444	holster	0.058000	exposed_private_parts	n03527444_12661.JPEG
n03617480	kimono	0.091925	exposed_private_parts	n03617480_6206.JPEG
n03710637	maillot	0.801976	beach_voyeur	ILSVRC2012_val_00021081.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_15836.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_272.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_3832.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_5095.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_5373.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_5386.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_66.JPEG
n03710637	maillot	0.801976	beach_voyeur	n03710637_7074.JPEG
n03710637	maillot	0.801976	exposed_private_parts	n03710637_6756.JPEG
n03710721	maillot, tank suit	0.768278	beach_voyeur	n03710721_1812.JPEG
n03710721	maillot, tank suit	0.768278	beach_voyeur	n03710721_3040.JPEG
n03710721	maillot, tank suit	0.768278	beach_voyeur	n03710721_3488.JPEG
n03710721	maillot, tank suit	0.768278	beach_voyeur	n03710721_7542.JPEG
n03710721	maillot, tank suit	0.768278	beach_voyeur	n03710721_8122.JPEG
n03710721	maillot, tank suit	0.768278	verifiably_pornographic	n03710721_25886.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_10283.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_18237.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_2462.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_2920.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_3615.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_4096.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_4203.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_6214.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_8550.JPEG
n03770439	miniskirt, mini	0.619425	upskirt	n03770439_9676.JPEG
n03770439	miniskirt, mini	0.619425	verifiably_pornographic	n03770439_12003.JPEG
n03770439	miniskirt, mini	0.619425	verifiably_pornographic	n03770439_1347.JPEG
n04209133	shower cap	0.130216	exposed_private_parts	n04209133_10606.JPEG
n04209133	shower cap	0.130216	exposed_private_parts	n04209133_206.JPEG
n04209133	shower cap	0.130216	exposed_private_parts	n04209133_716.JPEG

Table 5: Table containing the results of hand surveyed images

Reviewer-ID	Review
A- Grad student, CMU SCS	This one reminds me of a mix between graffiti and paper mache using newspaper with color images or magazines . My attention is immediately drawn to near the top of the image which, at first glance, appears to be a red halo of sorts, but upon further consideration, looks to be long black branching horns on a glowing red background. My attention then went to the center top portion, where the "horns" were coming from, which appeared to be the head or skull of a moose or something similar. The body of the creature appears to be of human-like form in a crucifix position, of sorts. The image appears more and more chaotic the further down one looks.
B- Grad student, Stanford CS	Antisymmetric: left side is very artistic, rich in flavor and shades; right is more monotonic but has more texture. Reminds me of the two different sides of the brain through the anti-symmetry
C- Data Scientist, Facebook Inc	Futurism
D- CS undergrad, U-Michigan	It's visually confusing in the sense that I couldn't tell if I was looking at a 3D object with a colorful background or a painting. It's not just abstract, but also mysteriously detailed in areas to the point that I doubt that a human created these
E - Senior software engineer, Mt View	The symmetry implies a sort of intentionally. I get a sense of Picasso mixed with Frieda Callo[sic] here.
F- Data Scientist, SF	Reminds me of a bee and very colorful flowers, but with some nightmarish masks hidden in some places. Very tropical

Table 6: Responses received for the neural art image in Fig 10