Issues in Computer Vision Data Collection: Bias, Consent, and Label Taxonomy

by

Chris Dulhanty

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contribution

Chapter 2 of this thesis contains content from the previously published paper:

> Chris Dulhanty and Alexander Wong. 2019. Auditing ImageNet: Towards A Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets. In *Workshop on Fairness Accountability Transparency and Ethics in Computer Vision (FATE/CV) at CVPR 2019, June 16-20, 2019, Long Beach, CA, USA.*

Alexander Wong contributed to the conceptualization of this work and to the editing of a completed draft of this paper.

Chapter 3 of this thesis appears, with minor edits, in its entirety as it does in the previously published paper:

> Chris Dulhanty and Alexander Wong. 2020. Investigating the Impact of Inclusion in Face Recognition Training Data on Individual Face Identification. In *2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES'20), February 7–8, 2020, New York, NY, USA.*

Alexander Wong contributed to the conceptualization of this work, to the editing of a completed draft of this paper, and to assisting in addressing reviewers' comments.

Abstract

Recent success of the convolutional neural network in image classification has pushed the computer vision community towards data-rich methods of deep learning. As a consequence of this shift, the data collection process has had to adapt, becoming increasingly automated and efficient to satisfy algorithms that require massive amounts of data. In the push for more data, however, careful consideration into decisions and assumptions in the data collection process have been neglected. Likewise, users accept datasets and their embedded assumptions at face-value, employing them in theory and application papers without scrutiny. As a result, undesirable biases, non-consensual data collection, and inappropriate label taxonomies are rife in computer vision datasets. This work aims to explore issues of bias, consent, and label taxonomy in computer vision through novel investigations into widely-used datasets in image classification, face recognition, and facial expression recognition. Through this work, I aim to challenge researchers to reconsider normative data collection and use practices such that computer vision systems can be developed in a more thoughtful and responsible manner.

Acknowledgements

I would like to thank my supervisors Alexander Wong and David Clausi for providing an incredible amount of support over the past two years, always encouraging me to pursue research questions that I was truly passionate about. The environment that you, Paul Fieguth, and John Zelek have fostered in the Vision and Image Processing Lab is unique, one that encourages intellectual curiosity and collaboration. To my lab mates, thank you for your guidance and partnership in coursework, research, and in navigating the ups and downs of graduate student life.

Thank you to my mom Cathy and dad Steve for your unwavering encouragement. The strong work-ethic and curiosity you instilled in me are qualities that I hold dear to this day. To my brother Kevin and sister and editor Emily, thank you for listening to my rants on this and other topics that you may or may not have found as interesting as I did. To my friends, thank you for your support over the past two years, I am indebted to you all. To Alexander and Andrew, thank you for our countless discussions on the societal impact of artificial intelligence, your push for me to pursue research in this area, and your teamwork in our hackathon debut in December 2018. This thesis is a direct result of my experience at that event. And to Heather, thank you for being an amazing partner, and for all your love and support.

Thank you to the University of Waterloo, the Vector Institute, and the National Research Council of Canada for supporting me with scholarships and programming to grow my knowledge and professional network. To Graham Taylor, thank you for allowing me to take your graduate course in machine learning in the fall of 2017, it was a foundational experience that solidified my desire to pursue a career in this field. Thank you to my thesis reading committee, comprising of my supervisors, Jennifer Boger, and Scott Campbell, for your time and feedback in preparing this work for publication, and to the reviewers of previously published versions of Chapters 2 and 3. Lastly, thank you to my colleagues in the fields of machine learning, computer vision, and fairness, accountability, transparency, and ethics in computing. I have learned an incredible amount from reading your work, listening to your talks, and on occasion, having the opportunity to discuss my research with you. This thesis stands on the shoulders of your incredible work.

Dedication

This thesis is dedicated to all those who fight for justice in the world. Black Lives Matter.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

> Computational depth without historic or sociological depth is superficial learning.
>
> — Ruha Benjamin, *Keynote at ICLR 2020* [9]

1.1 On the Shoulders of ImageNet

Data underpins deep learning. Deep learning, by way of convolutional neural networks (CNNs), dominates modern computer vision research and applications. Since Krizhevsky et al. demonstrated in 2012 that high-capacity CNNs trained with large amounts of data on graphical processing units result in powerful image classification systems [72], the field of computer vision has widely adopted this paradigm. Deep learning has been applied to medical image segmentation [115], human pose estimation [132], face recognition [128], and many more tasks with great success, establishing the CNN as the preeminent method in computer vision. As progress in the research community has historically been measured by accuracy on benchmark datasets such as MNIST [11], ImageNet [27], and COCO [79], researchers are motivated to collect more data, use more computing power, and train higher-capacity CNNs to further the state-of-the-art in their respective domains. Individuals are not alone in this pursuit, however, as a tenet of the research community is the sharing of data. As computer vision systems advance from handcrafted feature-based methods to supervised deep learning, and now to weak-supervision [125, 86] and self-training [143], increasingly large datasets are in demand. To this end, data collection practices in the computer vision community have shifted dramatically in the past thirty years.

In the 1990s, datasets were collected by academics in laboratory settings [100, 85] and made available through partnerships with government agencies [11, 106]. As the consumer internet boomed at the turn of the century, online search engines and social media websites provided a new means of collection. Computer vision researchers moved online in the 2000s to collect images, manually annotating datasets such as Caltech-101 [39] and PASCAL VOC [36], but these datasets, on the order of tens of thousands of examples, pushed the limit of in-house annotation. Fortunately for researchers, Amazon Mechanical Turk (AMT)[1], a website to hire remote crowdworkers to perform short, on-demand tasks, launched in 2005 and provided a solution. From 2007 to 2010, researchers at Stanford and Princeton used the crowdsourcing platform to task 49k "turkers" from 167 countries with annotating images to create the canonical ImageNet dataset of 14M images in 22k classes [38]. The associated ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was held annually from 2010 to 2017, created from a subset of 1k classes from the larger dataset. The research community coalesced around the ILSVRC as it presented an image classification problem an order of magnitude more difficult than its predecessors. Landmark work from Krizhevsky et al. [72] in the 2012 event was the catalyst for significant academic, industry, and state interest and investment in deep learning and artificial intelligence.

ImageNet and its challenge have been featured in the New York Times [87], cited more than 38k times [27, 117], and described as "the data that transformed AI research — and possibly the world." [47] ImageNet's success entrenched the use of crowdsourced annotations in the data collection pipeline, effectively solving the problem of large-scale data collection for the research community. This technique was subsequently used in collecting the object detection, segmentation and captioning dataset COCO [79], the human action classification dataset Kinetics [67], and the densely-annotated scene understanding dataset Visual Genome [71], among other widely-used datasets. The shift to web-scraped data and crowdsourced annotations, however, has not been without consequence.

1.2 The Abstraction of Data in Computer Vision

In the push for larger datasets to satisfy deep learning algorithms, careful considerations into the choices and assumptions underpinning data collection have largely been neglected. This is especially troubling in computer vision research, as many problem areas include the collection and interpretation images of human subjects, which brings with it issues of identity, privacy, and connotations of harmful classification systems from the past. The

[1] https://www.mturk.com/

automation of data sourcing and annotation emphasizes efficiency and indiscriminate collection above scrutiny. As Jo and Gebru write, "Taking data in masses, without critiquing its origin, motivation, platform and potential impact results in minimally supervised data collection" [64]. The distributed workforce accessed through AMT and similar platforms is often treated as a homogeneous, interchangeable group of annotators, ignoring cultural differences that can lead to different labels from different groups. Further, publications announcing datasets seldom provide rationale for the many value-laden decisions that were made in their composition [44], such as classification taxonomy and hierarchy, data source selection and representation, annotator instructions, compensation and demographics, and many more. These decisions embed biases and assumptions into data but are largely ignored as the focus of the community is on the product of data collection, not the process. While abstraction, the process of reducing complexity by considering something independent of its details [19], is a powerful concept in computer science, abstracting the social context away from data collection removes important details that are crucial to understand how a dataset represents the world.

In a similar vein, published datasets are largely accepted by practitioners for use in theory and application papers without scrutiny. With repeated use, datasets become viewed as neural scientific objects, the many subjective decisions that went into their construction rarely contested by the community [43]. For many in computer vision, the actual images in ImageNet have been abstracted away, replaced with a testing suite that evaluates the performance of an image classification model on the command line. The notion that the 1k classes in the widely-used 2012 ILSVRC subset of ImageNet are well-selected to act as the gold-standard benchmark for image classification is in itself an assumption that is uncontested.

1.3 Issues with Current Data Practices

The lack of rigour in the collection and the lack of scrutiny in the use of datasets in computer vision lead to consequences that are far-reaching.

1.3.1 Bias

Undesirable biases are patterns or behaviours learned from data that are highly influential in the decisions of a model, but not aligned with the values (or idealized values) of the society in which the model operates [124]. Bias in models arise from many different sources

in the machine learning development process and can occur with respect to age, gender, race, or the intersections of these and other protected attributes [126]. One source of bias occurs when training data underrepresents some subset of the population that the model sees as input when it is deployed. Many face recognition datasets have been shown to display this so-called *representational bias*, leading to poor performance of derived models on Black people, specifically Black women [13, 108, 91]. Such bias is very concerning as face recognition models are actively used by law enforcement agencies across the world, with reports of false positive identifications leading to wrongful arrests, as was the case with Robert Williams by Detroit police in January 2020 [55]. Likewise, some state-of-the-art object detection systems have been demonstrated to have worse performance in identifying pedestrians with darker skin tones [138]. As many autonomous driving companies rely on CNNs for visual understanding of the world, these reports are concerning.

Efforts to gather more diverse data to increase representation can prove difficult, however, as entrenched inequalities in society are often present at the source of collection. *Historical bias* is another means by which undesirable bias can be embedded in a computer vision system, as this bias exists given perfect sampling of the data source, a consequence of deep-rooted systemic unfairness [126]. Labeled Faces in the Wild [60], for example, is a gold standard benchmark in face verification [152]. It was sourced through images and captions of notable people in *Yahoo! News* stories from 2002 to 2004 and was estimated to contain 77.5% male and 83.5% white individuals [50]. This highly-skewed representation is a result of a Western-focused media source that brings with it a patriarchy and history of systemic racism that undervalues women and people of colour in leadership positions in business, politics, academia, entertainment, and other newsworthy professions. The decision to select identities in this manner embedded a historical bias in the dataset, of which no steps were taken to mitigate.

Bias can manifest in many other areas of the machine learning development process, as demonstrated in Figure 1.1. For a thorough conceptual framework for understanding bias, refer to work by Suresh and Guttag [126].

1.3.2 Consent

Consent and privacy are notions not well addressed by computer vision practitioners in web-based data collection. In Canada, research involving human subjects is exempt from Research Ethics Board review when it "relies on information that is in the public domain and the individuals to whom the information refers have no reasonable expectation of privacy" [14]. But to what extent do individuals give up their privacy expectation when they

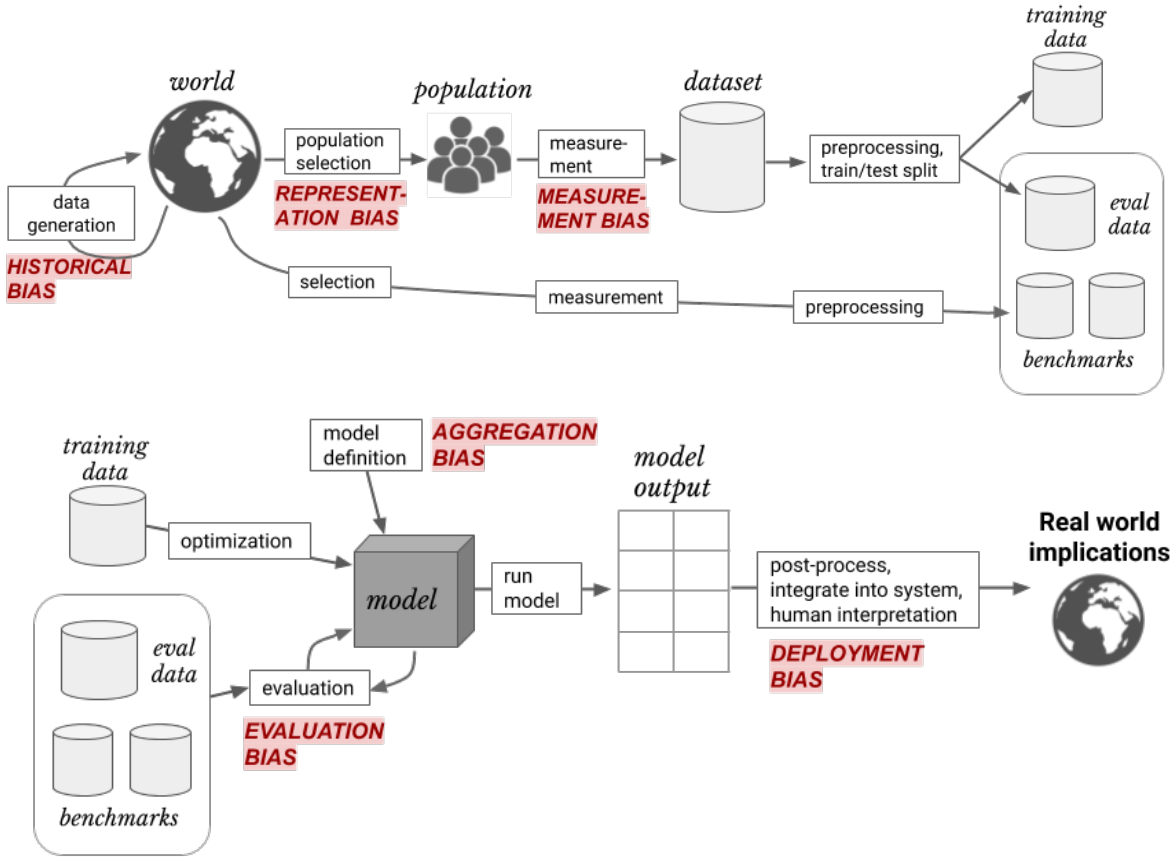

Figure 1.1: A framework for understanding sources of bias in machine learning development [126].

post content online, or when others post content of them without their consent? Critics of current research ethics regulations say the advent of big data dramatically changes the research ethics landscape, yet regulations have not been updated to address new challenges of web-based data collection [92]. Researchers often rationalize the collection of data in face recognition, for example, by restricting datasets to celebrity identities, as they view these individuals to have lower expectations of privacy, but this is not always the case. Some researchers provide a means for individuals, celebrity or not, to opt-out of inclusion in face datasets [2], signaling an appreciation of the non-consensual nature of their collection. However, the onus remains on the individual to uncover their inclusion in such datasets, which are often restricted to approved researchers.

In some jurisdictions however, individuals have legal protections against the non-consensual analysis of their face. The Biometric Information Privacy Act (BIPA) [61] is an Illinois State law enacted in 2008 that gives residents the right to seek financial compensation from private companies who conduct biometric analysis without obtaining informed consent, specifically mentioning face scans as a protected biometric. Potential financial liabilities of popular face recognition dataset MegaFace [69] were recently raised by legal experts in a New York Times exposé [56]. MegaFace, created by researchers at the University of Washington in 2016, was collected through publicly available images of non-celebrities on Flickr. It was taken offline in April 2020 [3]. Even as public discourse around data collection and consent increases, little effort has been displayed by computer vision researchers to engage with these issues in their work. As Solon writes in an NBC News report on the ethics of face recognition datasets, "It was difficult to find academics who would speak on the record about the origins of their training datasets; many have advanced their research using collections of images scraped from the web without explicit licensing or informed consent" [122].

1.3.3 Label Taxonomy

Labels in a dataset are often referred to as "ground truth" [117, 79, 71], yet this terminology often provides a veil of objectivity for annotations that are stereotypical, subjective, and lack scientific foundations linking them to images.

[2] https://github.com/NVlabs/ffhq-dataset/, "we are committed to protecting the privacy of individuals who do not wish their photos to be included."; https://web.archive.org/web/20180218212120/http://www.msceleb.org/download/sampleset, "Please contact us if you are a celebrity but do not want to be included in this data set. We will remove related entries by request"

[3] http://megaface.cs.washington.edu/

Datasets frame problems through their classification schema. ImageNet draws its taxonomy from WordNet [94], a lexical database developed in the 1990s at Princeton that organizes sets of synonyms, or "synsets", into semantically meaningful relationships, each expressing a distinct concept. During ImageNet's construction, 80k noun synsets were filtered through algorithmic and manual methods to arrive at 22k classes [27]. However, the extent to which each class can be characterized visually varies considerably. While a `football player` or `scuba diver` evoke clear visual pictures, a `stakeholder` or `hobbyist` cannot. Worse, some classes in ImageNet such as a `debtor`, `snob`, and `good person` promote stereotypes and ideas of physiognomy, the pseudoscientific assertion that one's personal essential character can be gathered from their outer appearance [23]. While ImageNet authors have made recent attempts to rectify this situation by removing explicitly offensive and non-imageable classes and diversifying others, this work comes more than ten years after the dataset's release and widespread use in the research community [145].

Annotations such as those in facial attractiveness dataset SCUT-FBP, which assigns an attractiveness label between one and five to 500 Asian women [142], prove problematic as they launder subjectivity through data. While research suggests some elements of faces such as facial symmetry are found universally attractive, perhaps as an evolutionary indicator of good health [80], this is far from absolute. The notion of beauty varies in time, geography, culture, and between individuals, so any attempt to create annotations that are treated as "ground truth" in perpetuity is fraught. The authors' attempt to mitigate this subjectivity by averaging results from several annotators speaks to the fundamental uncertainty in the annotation task. Taxonomy issues notwithstanding, the inclusion of only women in the SCUT-FBP dataset promotes objectification, especially considering only 13% of subjects in the database were captured by the researchers themselves, the remainder collected from the web without consent.

Physiognomy appears again in work by Wu and Zhang [141], entitled *Automated Inference on Criminality using Face Images*, in which face images are annotated as criminals and non-criminals in order to automate their identification with deep learning. An unwritten assumption in this work is that criminality is an innate class of individuals, linked to genetics, that manifests in the face. This line of thinking discounts an entire body of behavioural and social sciences that examines how socioeconomic status, lived experiences, environment, and other factors may impact criminality [78]. While the technical claims of this study have been rebuked [144], bigger questions arise with respect to the motivations and ethical implications of this work. Although the authors of this study claim their work is "only intended for pure academic discussions" and motivated by a curiosity of the visual capabilities of machine learning systems, such statements promote a problematic "view from nowhere" that discounts the world in which their research exists and power

imbalances therein, a perspective of scientific objectivity thoroughly critiqued by feminist scholars [51]. As companies such as Faception[4] claim the ability to identify terrorists and pedophiles from face images, research that is earnestly conducted out of curiosity can embolden commercialization and perpetuate harm, which is not evenly distributed in our unjust world, especially when used by a law enforcement establishment with a history of systemic racism [95, 89]. While an egregious case of a lack of research into domain-specific literature, this study is emblematic of a larger problem with data annotation that can uphold a visual relationship between an image and its label that is not founded in science.

1.4 Thesis Overview

This work aims to explore issues of bias, consent, and label taxonomy in computer vision through novel investigations into widely-used datasets in image classification, face recognition, and facial expression recognition. Through this work, I aim to challenge researchers to reconsider normative data collection and use practices such that computer vision systems can be developed in a more thoughtful and responsible manner.

1.4.1 Motivation

ImageNet [27] ushered in a flood of academic, industry, and state interest in deep learning and artificial intelligence. Despite ImageNet's significance, in the ten years following its publication at the leading computer vision conference CVPR in 2009, there was never a comprehensive investigation into the demographics of the human subjects contained within the dataset. This is concerning from a pragmatic perspective, as models trained on ImageNet are widely used by computer vision practitioners in transfer learning, the practice of applying knowledge acquired in one task to a different, but related problem. If certain groups are underrepresented in ImageNet, downstream models may inherit a biased understanding of the world. The extent to which possible biases are retained in models when trained on new datasets is an open question that cannot be answered until ImageNet is well-understood. From a cultural perspective, the lack of scrutiny into ImageNet is a prime example of how datasets are uncontested after publication. ImageNet has been championed as one of the most important breakthroughs in artificial intelligence and its achievements should indeed be celebrated, however it appears that either its success or a complacency in researchers lead to it not being studied critically for more than ten years,

[4] https://www.faception.com/our-technology

both of which are cause for concern. With this motivation I present Chapter 2 of this thesis, wherein I explore the question of bias in ImageNet by introducing a framework for the audit of large-scale image datasets.

With the advent of web-scraped data, informed consent in the collection of human subjects in face recognition datasets has been largely ignored. As such, modern datasets count in the millions of images and in the hundreds of thousands of identities. State-of-the-art face recognition systems leverage these large collections of *specific* individuals' faces to train CNNs to learn an embedding space that maps an *arbitrary* individual's face to a vector representation of their identity. The performance of a face recognition system is directly related to the ability of its embedding space to discriminate between identities, ergo, the size of its dataset. Recently, there has been significant public scrutiny into the source and privacy implications of large-scale face recognition datasets such as MS-Celeb-1M and MegaFace [52, 122, 56]. In 2005, an image of five-year-old Chloe Papa was uploaded to Flickr by their mother. In 2016, it was scraped and included in MegaFace. In 2019, Papa said to the New York Times regarding their inclusion, "It's gross and uncomfortable, I think artificial intelligence is cool and I want it to be smarter, but generally you ask people to participate in research. I learned that in high school biology" [56]. Many people are uncomfortable with their face being used to advance dual-use technologies such as face recognition that can enable mass surveillance. But is there a demonstrated impact of being included in such datasets? Are those included in the training sets of face recognition systems at a higher likelihood of being identified? This question has not previously been studied. In Chapter 3 of this thesis, I conduct experiments on a state-of-the-art face recognition system in an attempt to answer this question and further the discussion of privacy and consent in the context of data collection.

Facial expression recognition aims to predict the emotion a person is experiencing by analyzing images of their face. Research in this domain is built upon the work on Paul Ekman, a psychologist and researcher who has studied the relationship between emotions and facial expressions for more than 60 years. Ekman contends that a person's emotional state can be readily inferred from their face due to the universality of six basic emotions, consistent across cultures and individuals [34]. A landmark review study published in July 2019, however, says otherwise [7]. The review, spearheaded by psychologist Lisa Feldman Barrett, analyzed over 1,000 research papers that studied healthy adults across cultures, newborns and young children, and people who are congenitally blind to determine the reliability and specificity of facial expressions in identifying emotions. Their findings vehemently refute Ekman's claims. When interacting with others, we do not just rely on their face to try to infer their emotional state, but body language, tone of voice, word choice, situational context, our relationship, cultural norms, and other factors contribute

to our ability to do so. In its current problem formulation, facial expression recognition with computer vision abstracts all of this context away, reducing the complex task to a classification problem with static images. While people may smile when happy, the use of the label "happy" on a static image of a grinning face does not have a solid scientific foundation. With firms such as HireVue using facial expression recognition to screen candidates in job interviews [53], continued research in its current form can bolster unproven technologies that have considerable impacts in people's lives. In Chapter 4 of this thesis, I revisit the canonical Japanese Female Facial Expression (JAFFE) dataset, widely used in facial expression recognition research, and analyze its collection and use in the context of the aforementioned review, in the hopes of communicating these findings to a larger audience.

1.4.2 Contributions

The main contributions of this thesis include: the introduction of a model-driven demographic annotation pipeline for apparent age and gender in large-scale image datasets; the presentation and analysis of the first audit of the 2012 ILSVRC subset of ImageNet (1.28M images) and the `person` subtree of ImageNet (1.18M images); the first evidence of differential identification accuracy for individuals by a state-of-the-art face recognition system in one-to-many searches, dependent on their inclusion or exclusion in training data; and a novel analysis of the JAFFE dataset through the lens of Barrett et al.'s [7] criteria for identifying emotional inference from facial expression images.

1.4.3 Outline

The remainder of this thesis is organized in the following manner. Chapter 2, entitled "Bias in Large-Scale Image Datasets: ImageNet Demographics Audit," details experiments auditing ImageNet for bias on the axes of gender and age. Chapter 3, entitled "Consent in Face Recognition: Impact of Individual Inclusion in Training Data," outlines the history of web-scraped data in face recognition research and details experiments assessing the impact on an individual's inclusion in such datasets. Chapter 4, entitled "Label Taxonomy in Facial Expression Recognition: Recontextualizing the Problem," investigates issues in the formulation of facial expression recognition research through the analysis of the JAFFE dataset. This thesis concludes with Chapter 5, were I bring together ideas learned from the three experiments and discuss initiatives to move the computer vision community forward in a positive direction.

Chapter 2

Bias in Large-Scale Image Datasets: ImageNet Demographics Audit

2.1 Introduction

ImageNet [27], released in 2009, is a canonical dataset in computer vision. ImageNet follows the WordNet lexical database [94], which groups words into "synsets," each expressing a distinct concept. ImageNet contains 14,197,122 images in 21,841 hierarchical synsets, collected through a comprehensive web-based search and annotated with Amazon Mechanical Turk [27]. An example of images in ImageNet's hierarchical structure is seen in Figure 2.1. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [117], held annually from 2010 to 2017, was the catalyst for an explosion of academic, industry, and state interest in deep learning. A subset of 1,000 non-overlapping internal and leaf nodes of ImageNet were used in the ILSVRC classification task; seminal work by Krizhevsky et al. in the 2012 event cemented the convolutional neural network (CNN) as the preeminent model in computer vision [72].

Today, work in computer vision largely follows a standard process: a pretrained CNN is downloaded with weights initialized to those trained on the 2012 ILSVRC subset of ImageNet (ILSVRC-2012), the network is adjusted to fit the desired task, and transfer learning is performed, whereby the CNN uses the pretrained weights as a starting point for training new data on the new task. The use of pretrained CNNs is instrumental in applications as varied as remote sensing [88], cervical cell classification [150], and chest radiograph diagnosis [109], for example.

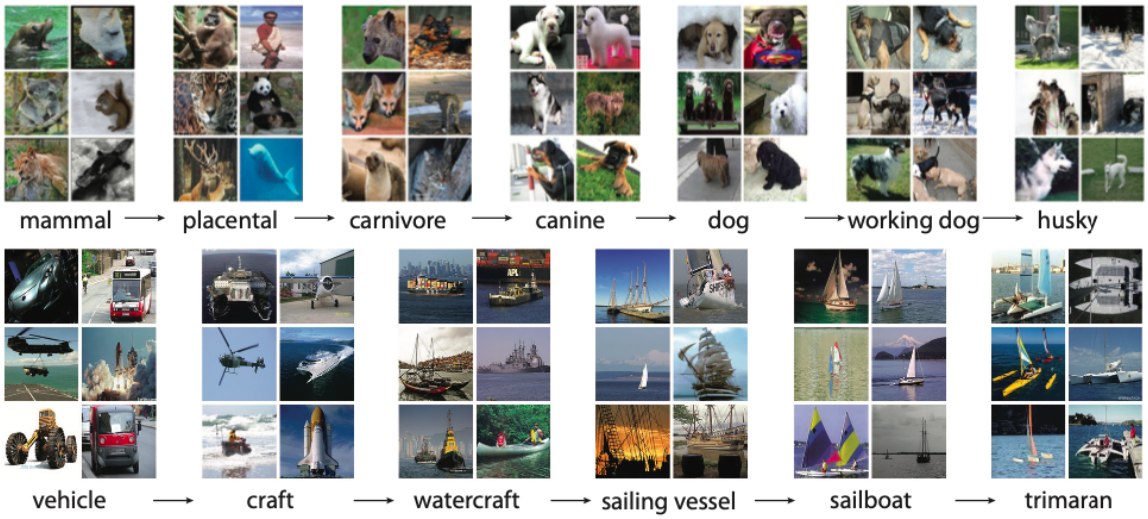

Figure 2.1: A snapshot of two root-to-leaf branches of ImageNet. For each synset, six randomly sampled images are presented [26].

By convention, computer vision practitioners have effectively abstracted away the details of ImageNet. While this has proven successful in practical applications, there is merit in taking a step back and scrutinizing common practices. In ten years following the release of ImageNet, a comprehensive study into the composition of images in the classes it contains was never conducted.[1] The lack of scrutiny into ImageNet's contents is concerning, given evidence from Roberts that CNNs trained on ILSVRC-2012 implicitly encode age, gender, and race information and can pass this information to derivative models during transfer learning [114]. As Roberts writes on the implications of her 2018 thesis,

> Are researchers, hobbyists, and government agencies using and adapting these and similar models for their visual classification tasks without fully understanding how features, unbeknownst to them (and especially features that encode protected attributes), are indirectly influencing the outcome of their results (and in potentially biased and harmful ways)? [114]

Without a conscious effort to incorporate diversity in data collection, undesirable biases can collect and propagate. Age, gender, and racial biases have been identified in word

[1]Since the publication of a preliminary version of this chapter in June 2019 [33], several studies have also completed audits of ImageNet [145, 107]. These works are discussed in the following section.

embeddings [10], image captioning models [3], and commercial computer vision gender classifiers [13], the result of biased data. The extent to which such biases exist in ImageNet is the topic of this chapter.

With this context, this chapter aims to address the research question: what is the demographic distribution of the training set of ILSVRC-2012 (1.28M images) and the `person` hierarchical synset of ImageNet (1.18M images), with respect to apparent age and gender presentation?

2.2 Analysis of ImageNet

In 2017, Shankar et al. studied the geo-diversity of ImageNet by analyzing metadata of the 14M images contained in the dataset [119]. They found a majority of images were sourced from North America and Western Europe, 45% originating from the US, while Indian and Chinese-based images accounted for 2.1% and 1% of the dataset, respectively. In 2018, Stock and Cisse introduced a novel tool for uncovering biases learned by models, applying it to CNNs trained on ILSVRC-2012 [124]. Using adversarial examples as a form of model criticism, they discovered that prototypical examples of the synset `basketball` contained images of Black people at a much higher rate than their white counterparts, despite a relative racial balance within the class. They hypothesized that an under-representation of Black people across all classes of ILSVRC-2012 may have lead to a biased representation of `basketball`.

ImageNet Roulette[2] was a provocation created by Crawford and Paglen that went viral in September 2019 [93, 5]. Users uploaded images of a person to the online art project and a CNN trained on the `person` subtree of ImageNet would classify them into one of 2,833 classes, as shown in Figure 2.2. The project and its associated essay *Excavating AI* sought to ignite a conversation on the politics of machine learning training data by demonstrating what happens when a system is trained on a problematic dataset [23]. Through the project, many overtly racist, misogynistic and offensive categories of ImageNet were uncovered. This work lead to an announcement[3] by the creators of ImageNet of upcoming work to filter and balance the `person` subtree. In this work, published at FAT* 2020, Yang et al. [145] identified 1,593 unsafe synsets either explicitly offensive or offensive depending on context, and 2,614 with a low "imageability" score, that is, abstract categories difficult to characterize accurately with images, such as `philanthropist`. This analysis left 158

[2] https://web.archive.org/web/20190926014608/https://imagenet-roulette.paglen.com/
[3] http://image-net.org/update-sep-17-2019

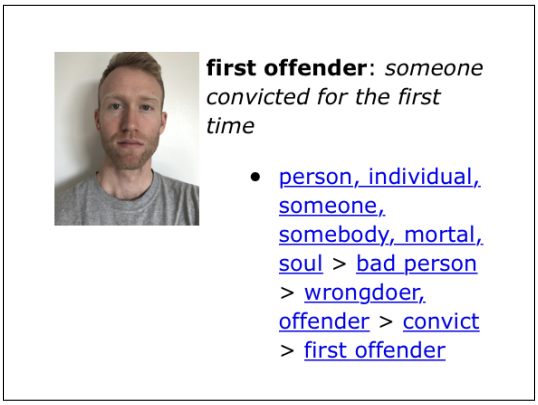

Figure 2.2: Example output of the *ImageNet Roulette* project.

synsets considered both safe and imageable, 139 of which contained at least 100 images that were balanced on the axes of skin colour, age, and gender [145]. Following this work, Prabhu and Birhane recently performed a comprehensive *ImageNet Census* of ILSVRC-2012, providing synset-level analysis across 61 metrics, and discussed many societal and ethical implications of large-scale data collection [107]. Among their findings were verifiable pornographic images in ILSVRC-2012, including images taken in non-consensual settings, such as up-skirts in the `miniskirt` synset.

2.3 Diversity Considerations in Creating ImageNet

Before proceeding with annotation, there is merit in contextualizing this study with a look at the methodology proposed by Deng et al. [27] in the construction of ImageNet. A close reading of their data collection and quality assurance processes demonstrates that the conscious inclusion of demographic diversity in ImageNet was lacking.

First, candidate images for each synset were sourced from commercial image search engines, including Google, Yahoo!, Microsoft's Live Search, Picsearch, and Flickr [37]. However, gender and racial biases have been demonstrated to exist in image search results, such as in the representation of people in images returned for the keyword "CEO" [66, 102]. As these studies and any resulting actions to mitigate their findings come several years after the collection of ImageNet, this demonstrates that a more curated approach at the top of the funnel may have been necessary to mitigate historical biases of search engines. Next, English search queries were translated into Chinese, Spanish, Dutch, and Italian

using WordNet databases and used for image retrieval. While this is a step in the right direction, Chinese was the only non-Western European language used. Other resources, such as Universal Multilingual WordNet, were available at the time, which included over 200 languages for translation [25]. Finally, the authors quantified the diversity of their dataset by computing the average image of each synset and measuring its lossless JPEG file size. They stated that a diverse synset would result in a blurrier average image and a smaller file size, representative of diversity in appearance, position, viewpoint, and background. While quantifying diversity in the global appearance of images, this method did not quantify diversity with respect to demographic characteristics such as age, gender, and Fitzpatrick skin type [40].

2.4 Methodology

In order to provide demographic annotations at scale, there exist two feasible methods: crowdsourcing and model-based annotations. In the case of large-scale image datasets, crowdsourcing quickly becomes prohibitively expensive. Model-based annotations use supervised learning methods to create models that can predict annotations, but this approach comes with its own meta-problem: as the goal of this work is to identify the representative demographics in a dataset, the annotation models must first be analyzed for performance on intersectional groups to determine if they exhibit bias themselves.

2.4.1 Criteria for Bias and Poor Representation

Age and gender groups are defined to have an approximately equal population in each intersectional group, according to US Census Bureau data from 2009[4], the year of ImageNet's release. These population-level statistics are displayed in Table 2.1. An American population base was selected as Shankar et al. [119] found US-based images to be the largest cohort in ImageNet. In reporting results of annotation models, a bias is said to exist if the performance on a group is 10% worse than the best-performing group along the axis or axes in question. In reporting results of representation in ImageNet, a group is said to be over or under-represented if it differs by more than 10% from its population-level statistic. It should be noted that some synsets of ImageNet have inherent gender or age connotations, as with specific clothing types, such as `bikini`, or in descriptions of people, such as `senior`. The focus of this work is on synsets in which undesirable biases may exist.

[4] https://www.census.gov/data/tables/2009/demo/age-and-sex/2009-age-sex-composition.html

% of US Population			
	Gender		
Age Group	**Male**	**Female**	**All**
0-14	10.41	9.94	20.35
15-29	10.64	0.30	20.93
30-44	9.97	10.12	20.10
45-59	10.23	10.70	20.94
60+	7.87	9.82	17.69
All	49.12	50.88	100.00

Table 2.1: US Census Bureau: Population Statistics by Age and Sex for 2009.

2.4.2 Face Detection

The FaceBoxes network [151] is employed for face detection, consisting of a lightweight CNN that incorporates novel Rapidly Digested Convolutional Layers and Multiple Scale Convolutional Layers for speed and accuracy, respectively. This model was trained on the WIDER FACE dataset [146] and achieves average precision of 95.50% on the Face Detection Data Set and Benchmark (FDDB) [62]. On a subset of 1,000 images from FDDB hand-annotated by the author for apparent age and gender presentation, the model achieves a relatively equal performance across intersectional groups, as show in Table 2.2.

Average Precision (%)			
	Gender Presentation		
Age Group	**Masculine**	**Feminine**	**All**
0-14	93.75	100.00	97.44
15-29	99.20	100.00	99.61
30-44	100.00	99.11	99.76
45-59	100.00	100.00	100.00
60+	99.17	100.00	99.24
All	99.48	99.74	99.54

Table 2.2: Face detection model average precision on a subset of FDDB, hand-annotated by the author for apparent age and gender presentation. Biased groups in bold.

2.4.3 Apparent Age Annotation

The task of apparent age annotation arises as ground-truth ages of individuals in web-scraped images are not possible to obtain. This work follows Merler et al. [91] and employs the Deep EXpectation (DEX) model of apparent age [116], pretrained on the IMDB-WIKI dataset of 500k faces with real ages and fine-tuned on the APPA-REAL training and validation sets of 3.6k faces with apparent ages, crowdsourced from an average of 38 votes per image [2]. As show in Table 2.3, the model achieves mean average error of 5.22 years on the APPA-REAL test set, but exhibits worse performance on younger and older groups.

Mean Average Error (years)			
	Gender		
Age Group	**Male**	**Female**	**All**
0-14	**8.26**	**9.37**	**8.77**
15-29	3.24	3.65	3.57
30-44	**4.52**	**4.93**	**4.72**
45-59	**4.43**	**5.50**	**4.81**
60+	**7.70**	**9.48**	**8.40**
All	5.13	5.29	5.22

Table 2.3: Apparent age model mean average error in years on APPA-REAL test set. Biased groups in bold.

2.4.4 Gender Presentation Annotation

Automated gender recognition (AGR) from face images is scientifically flawed [42]. Privacy risks and potential harms can result from individuals being incorrectly gendered or misgendered by such systems [49]. AGR perpetuates the view of a binary gender that excludes trans and gender-nonconforming identities [70]. The use of a model-based gender annotation tool in this work is done with the goal of identifying systematic bias in ImageNet in the hopes of improving inclusively, but I recognize such a system upholds gender normative expectations of presentation.

In this work, the gender variable of a detected face is expressed as a continuous value between 0 and 1. The labels `male` and `female` are used under the category `gender` to align with evaluation benchmarks that use this taxonomy. In reporting results on ImageNet, labels `masculine` and `feminine` are used under the category `gender presentation`. In both cases, the gender variable is thresholded at 0.5 to delineate classes. Following Merler

et al. [91], a DEX model is employed for these tasks. When tested on APPA-REAL, with enhanced annotations provided by Clapés et al. [17], the model achieves an accuracy of 91.00%, however its errors are not evenly distributed, as shown in Table 2.4. The model errs more on younger and older age groups and on those with a female gender label.

Accuracy (%)			
	Gender		
Age Group	**Male**	**Female**	**All**
0-14	**79.19**	**75.78**	**77.62**
15-29	96.09	89.80	91.95
30-44	100.00	91.72	95.99
45-59	100.00	91.30	96.89
60+	**84.91**	**79.71**	**82.86**
All	94.15	88.04	91.00

Table 2.4: Gender model binary classification accuracy on APPA-REAL test set. Biased groups in bold.

Further evaluation of the model is conducted on the Pilot Parliaments Benchmark (PPB) [13], a face dataset developed by Buolamwini and Gebru for parity in gender and Fitzpatrick skin type. Results for intersectional groups on PPB are shown in Table 2.5. The model performs poorly for darker-skinned females (Fitzpatrick skin types IV - VI), with an average accuracy of 69.00%, reflecting the disparate findings of commercial computer vision gender classifiers in *Gender Shades* [13]. The use of this model in annotating ImageNet will result in biased gender presentation annotations, under-representing feminine presenting identities to some extent, but I proceed to establish a baseline upon which a more fair model can improve annotations in future work.

2.5 Results and Discussion

The proposed methodology is applied to the training set of ILSVRC-2012, comprised of 1,000 synsets containing between 732 and 1,300 images, and the `person` subtree of ImageNet, comprised of 2,833 synsets containing between 1 and 1664 images. Face detections that receive a confidence score of 0.9 or higher move forward to the annotation phase. Statistics for gender presentation and age demographics of ILSVRC-2012 and the `person` subtree of ImageNet are presented in Table 2.6 and Table 2.7, respectively.

Accuracy (%)			
	Gender		
Fitzpatrick Skin Type	**Male**	**Female**	**All**
I	100.00	98.31	99.12
II	100.00	97.14	98.86
III	100.00	95.08	97.67
IV	100.00	**86.11**	92.31
V	100.00	**68.79**	**83.70**
VI	100.00	**62.77**	**86.22**
ALL	100.00	**83.57**	92.68

Table 2.5: Gender model binary classification accuracy on PPB. Biased groups in bold.

% of Dataset			
	Gender Presentation		
Age Group	**Masculine**	**Feminine**	**All**
0-14	4.06	3.01	7.08
15-29	**27.11**	**23.72**	**50.83**
30-44	**17.81**	9.02	**26.83**
45-59	8.48	5.08	13.56
60+	0.91	0.80	1.71
All	**58.38**	41.62	100.00

Table 2.6: Top-level age and gender presentation statistics of ILSVRC-2012. Over-represented groups in bold. Under-represented groups are underlined.

% of Dataset			
	Gender Presentation		
Age Group	**Masculine**	**Feminine**	**All**
0-14	3.35	1.66	5.00
15-29	**24.63**	**16.95**	**41.58**
30-44	**21.02**	7.36	**28.38**
45-59	16.87	3.70	20.57
60+	3.02	1.44	4.47
All	**68.89**	31.11	100.00

Table 2.7: Top-level age and gender presentation statistics of ImageNet `person` subtree. Over-represented groups in bold. Under-represented groups are underlined.

In these preliminary annotations, feminine presenting identities comprise 41.62% of images in ILSVRC-2012 and 31.11% of images in the `person` subtree of ImageNet. People who appear over the age of 60 are almost non-existent in ILSVRC-2012, accounting for 1.71% of all observed people.

There exist only three synsets in ILSVRC-2012 that are derived from the `person` subtree of ImageNet, `scuba diver`, `groom`, and `ballplayer`. Yang et al. use this fact to justify their work filtering and balancing only the `person` subtree of ImageNet, and not ILSVRC-2012 [145]. This work, however, finds 8.23% of all images in ILSVRC-2012 contain at least one person, and 29 synsets contain people in more than half of their images. The 20 synsets containing the most images with people are listed in Table 2.8, along with the percent of observed people in each synset that present feminine (percent presenting masculine is implied). Among these are synsets with connotations to science, academia, and law, which would be desirable to be balanced in terms of gender presentation. In synsets `lab coat, laboratory coat` and `academic gown, academic robe, judge's robe`, feminine presenting individuals comprise only 38.46% and 40.82% of observed people, respectively, demonstrating an under-representation in these categories with respect to the population-level statistics.

Note that our methodology is limited as it does not identify a person in an image if their face is not observed. Such a method will not identify bias with respect to the co-occurrence of certain skin types or body types with objects in images. Such a bias was recently observed in the Google Vision API, where the image of a hand with dark skin holding a thermometer was classified as a `gun`, while a version of the same image with light skin produced the label `monocular` [68].

To get a sense of the most skewed classes in terms of gender presentation for each dataset, synsets were filtered to retain those that contained at least 20 images and observed a person in at least 15% of their images. The percent of observed people with masculine or feminine presented identities in each synset were then calculated and ranked in descending order. The 20 most masculine presenting and feminine presenting synsets in ILSVRC-2012 are shown in Table 2.9 and Table 2.10, respectively. Top synsets for masculine presented identities largely represent types of fish, sporting implements, musical instruments, and firearm-related items, while top synsets for feminine presented identities largely represent types of clothing and dogs.

The 20 most masculine presenting and feminine presenting synsets in the `person` subtree of ImageNet are presented in Table 2.11 and Table 2.12, respectively. Contrary to the relatively benign synsets in ILSVRC-2012, this analysis surfaces many troubling classes related to people. Among the top synsets for masculine presented identities include harm-

Synset	% Containing People	% Presenting Feminine
lab coat, laboratory coat	76.23	38.46
wig	74.85	68.60
groom, bridegroom	74.15	42.09
pajama, pyjama, pj's, jammies	71.54	64.65
suit, suit of clothes	71.38	14.79
mortarboard	71.38	46.78
academic gown, academic robe, judge's robe	69.00	40.82
bow tie, bow-tie, bowtie	66.00	17.84
barracouta, snoek	65.69	5.76
ice lolly, lolly, lollipop, popsicle	65.38	55.40
bikini, two-piece	64.62	85.91
maillot	63.69	85.90
kimono	63.54	71.28
neck brace	63.14	53.96
military uniform	63.00	17.27
gown	62.00	83.62
fur coat	60.31	73.09
tench, Tinca tinca	59.54	7.60
seat belt, seatbelt	58.38	51.82
feather boa, boa	58.23	66.37

Table 2.8: ILSVRC-2012 synsets, by percent of images in a synset that contain at least one person.

ful categories such as `anti-American`, `enemy`, and `spree killer`, various classes related to business, including `oilman`, `pitchman`, `traveling salesman`, and `spellbinder`, promoting the stereotype of a male-centred vocation, and, as seen in ILSVRC-2012, sports-related synsets such as `second baseman` and `split end`. The top synsets for feminine presented identities contain a disturbing number of objectionable categories that have sexually-charged language, a phenomenon which is notably absent from the top masculine presenting synsets. Some of these synsets and their WordNet definitions include: `bombshell`, an entertainer who has a sensational effect; `tempter`, a person who tempts others; `nymph`, a voluptuously beautiful young woman; `nymphet`, a sexually attractive young woman; `smasher`, a very attractive or seductive looking woman; `wanton`, lewd or lascivious woman; `inguine`, an artless innocent young girl (especially as portrayed on the stage); `rosebud`, (a literary reference to) a pretty young girl. The sheer number of classes in this domain illustrates that the misogynistic taxonomy of WordNet was not adequately filtered out by ImageNet's creators during collection, likely due to the minimal supervision that came with outsourcing annotations to a distributed workforce.

Synset	% Presenting Masculine
barracouta, snoek	94.24
ballplayer, baseball player	93.47
Windsor tie	93.29
gar, garfish, garpike, billfish	92.76
tench, Tinca tinca	92.40
rugby ball	91.96
barbershop	90.06
bulletproof vest	89.33
sax, saxophone	89.16
swimming trunks, bathing trunks	88.36
assault rifle, assault gun	88.12
sturgeon	87.61
cornet, horn, trumpet, trump	87.15
trombone	86.68
suit, suit of clothes	85.21
football helmet	84.11
basketball	83.33
coho, cohoe, coho salmon, blue jack	82.80
military uniform	82.73
parallel bars, bars	82.63

Table 2.9: ILSVRC-2012 synsets, by percent of observed people in a synset presenting masculine.

2.6 Chapter Summary

The key takeaways of this chapter are:

- ImageNet is the gold-standard benchmark by which the computer vision community measures progress.
- Despite ImageNet's impact and widespread use, little attention has been paid to its demographic representation and label taxonomy in the ten years following its release, which is problematic as ImageNet is widely used to pretrain CNNs for transfer learning.
- Indeed, in analyzing the origins of ImageNet, it is clear that scale was prioritized over representation across age, gender, and racial lines in the collection process.

Synset	% Presenting Feminine
brassiere, bra, bandeau	88.59
golden retriever	88.29
bikini, two-piece	85.91
maillot	85.90
maillot, tank suit	85.47
miniskirt, mini	85.14
cocker spaniel, English cocker spaniel, cocker	83.82
gown	83.62
lipstick, lip rouge	82.81
Maltese dog, Maltese terrier, Maltese	82.06
stole	81.64
cardigan	81.32
overskirt	81.28
abaya	80.68
beagle	77.67
poncho	77.02
hoopskirt, crinoline	75.65
bonnet, poke bonnet	74.81
Labrador retriever	74.14
fur coat	73.09

Table 2.10: ILSVRC-2012 synsets, by percent of observed people in a synset presenting feminine.

- A model-based approach to identify the demographic representation of large-scale image datasets such as ImageNet is useful to mitigate high costs of manual annotation, however, models must first be audited for bias to ensure they perform fairly.
- Using a novel demographic annotation process, the ILSVRC-2012 dataset and the `person` subtree of ImageNet were found to contain a distinct lack of representation of young (0-14) and older (60+) aged people, and of feminine presenting individuals.
- This analysis uncovered a high representation of masculine presenting identities in academic, scientific, sporting, and business-related categories, and the inclusion of many categories of a sexually-charged nature that contained predominately feminine presenting identities.

ImageNet is emblematic of the emphasis on scale over scrutiny in the collection and

Synset	% Presenting Masculine
argonaut	100.00
agnostic	100.00
anti-American	100.00
chandler	100.00
counterterrorist	100.00
enemy	100.00
equerry	100.00
Girondist, Girondin	100.00
hacker	100.00
halberdier	100.00
helmsman, steersman, steerer	100.00
Kennan, George F. Kennan, George Frost Kennan	100.00
oilman	100.00
pitchman	100.00
second baseman	100.00
spree killer	100.00
shirtmaker	100.00
spellbinder	100.00
split end	100.00
traveling salesman	100.00

Table 2.11: ImageNet person subtree synsets, by percent of observed people in a synset presenting masculine.

use of data in the computer vision community. It is telling that a dataset that is so widely celebrated is also one that is rife with issues of poor representation and offensive label taxonomies, ignored by the research community for a decade as they focused more on getting their algorithms to work than on the societal impacts of their work [5]. As with Buolamwini, Gebru, and Raji in facial analysis systems in *Gender Shades* and its follow-up work [13, 108], it was not until Roberts [114], Shankar [119], and Cisse [124], underrepresented minority scholars in the computer vision community, thought to interrogate the assumptions of these widely uncontested systems that evidence began to arise with respect to biases embedded within them. In the case of ImageNet, the results of this work demonstrate that poor representation goes beyond the person subtree of ImageNet, but is also present in the commonly-used ILSVRC-2012 dataset. The extent to which biased feature representations are derived from ILSVRC-2012 and passed to downstream models during transfer learning remain open questions. What is certain from this work, however, is that

Synset	% Presenting Feminine
bombshell	100.00
choker	100.00
dyspeptic	100.00
comedienne	96.30
cover girl, pin-up, lovely	96.24
tempter	95.92
nymph, houri	95.74
artist's model, sitter	95.05
nymphet	93.84
smasher, stunner, knockout, beauty	93.84
maid, maiden	93.05
model, poser	93.01
wanton	92.70
lass, lassie, young girl, jeune fille	92.52
outdoorswoman	91.89
ingenue	91.85
frontierswoman	91.67
newswoman	91.67
rosebud	91.58
ingenue	91.30

Table 2.12: ImageNet person subtree synsets, by percent of observed people in a synset presenting feminine. Bold terms are

minimal supervision in collection and annotation can lead to datasets that are unequal across demographic axes. Datasets hold an enormous amount of power in the computer vision community, as they craft the problems that researchers organize themselves around. It is imperative that researchers attend more to issues of fairness and inclusion in the construction of datasets, as once they are released, it is incredibly difficult to effectively alter a dataset or change the norms of its use.

Chapter 3

Consent in Face Recognition: Impact of Individual Inclusion in Training Data

3.1 Introduction

Face recognition systems using CNNs depend on the collection of large image datasets containing thousands of sets of *specific* individuals' faces for training. Using this data, CNNs learn a set of parameters that can map an *arbitrary* individual's face to a feature representation, or *faceprint*, that has small intra-class and large inter-class variability. The ability of a face recognition system to distinguish between identities within this embedding space depends on the size and diversity of its training data, along with its model capacity and underlying algorithms. Face recognition systems have benefited from the enabling power of the Internet in the collection of large-scale image datasets and from hardware improvements in enabling efficient training of large models. Recently, increased attention to face recognition by academia, industry, and government has brought new researchers, ideas, and funding to the field, leading to performance improvements on benchmark tasks Labelled Faces in the Wild (LFW) [60] and MegaFace [99]. Consequently, face recognition systems are now being integrated into consumer and industrial electronic devices and offered as application programming interfaces (APIs) by providers such as Amazon, Microsoft, IBM, Megvii, and Kairos. However, along with improved performance has come increased public discourse on the ethics of face recognition development and deployment.

Algorithmic auditing of commercial face analysis applications has uncovered disparate

performance for intersectional groups across several tasks. Poor performance for darker skinned females by commercial face analysis APIs has been reported by Buolamwini, Gebru, and Raji [13, 108], as has lower accuracy in face identification by commercial systems with respect to lower (darker) skin reflectance by researchers at the US Department of Homeland Security [21]. As bias in training data begets bias in model performance, efforts to create more diverse datasets for these tasks have resulted. IBM's Diversity in Faces dataset [91], released in January 2019, is a direct response to this body of research. Using ten established coding schemes from scientific literature, researchers annotated one million face images in an effort to advance the study of fairness and accuracy in face recognition. However, this dataset has seen public scrutiny from a different, but equally notable perspective. A March 2019 investigation by NBC News into the origins of the dataset brought to the public conversation the issue of informed consent in large-scale academic image datasets, as IBM leveraged images from Flickr with a Creative Commons Licence without notifying content owners of their use [122].

To rationalize the collection of large-scale image datasets without explicit consent of individuals, some computer vision researchers appeal to the non-commercial nature of their work. However, work by Harvey and LaPlace at MegaPixels have found that authors' stated limitations on dataset use do not translate to real-world restrictions [52]. In the case of Microsoft's MS-Celeb-1M dataset, authors included an explicit "non-commercial research purpose only" clause with the dataset, which was the largest publicly-available face recognition dataset at the time. However, as the dataset has been cited in published works by the research arms of many commercial entities, findings cannot easily be isolated from improvements in product offerings. As a direct result of MegaPixel's work on the ethics, origins, and privacy implications of face recognition datasets, MS-Celeb-1M [48], Stanford's Brainwash dataset [123], and Duke's Multi-Target, Multi-Camera dataset [113] were removed from their authors' websites in June 2019. However, data remains accessible via torrents, derived datasets and other hosts [52].

In addition to issues of bias and informed consent in data collection, concern over the general use of face recognition systems by commercial and government agencies has been raised by civil rights groups and research centers, as there is no oversight for its use in civil society [1, 137]. For these and other reasons, multiple cities in the United States have banned the use of face recognition systems for law enforcement purposes [20, 140, 110]. Many people are concerned with their identify being used to train the dual-use technology that is face recognition. With reports of face recognition being used by law enforcement entities to identify protesters in Baltimore [111], London [12], and Hong Kong [96] there is merit in understanding the impact of one's inclusion in the training data that fuels the development of these systems.

With this context, this chapter aims to address the research question: is there a differential impact of face recognition systems in one-to-many search accuracy, dependant upon an individual's inclusion in the training dataset?

3.2 Background

3.2.1 Face Recognition Tasks

Within the domain of face recognition lies two categories of tasks: *face verification* and *face identification* [73].

In face verification, the goal is to assess if a presented image matches with the reference image of an individual, often to grant access to a physical device or location. Unlocking a smartphone with one's face provides an example of face verification; a person presents their face to a phone and it is verified against a reference image of the known owner of the device. This task is referred to as a one-to-one (1:1) search, as there is only one individual that the presented face image is compared against. In order to confirm a match, a threshold of similarity must be met, which can be set by the user of a system to meet a specific level of security. Performance of a system on face verification tasks is reported in terms of accuracy, the number of correct verifications of all verification attempts.

In face identification, a *gallery* of known identities is constructed from face images of individuals in advance of testing. Subsequently, a face image of unknown identity is presented to the system as the *probe*. The probe is then matched for similarity with all images in the gallery, constituting a one-to-many (1:N) search. If the system guarantees that the identity of the probe is within the gallery of identities, the problem is considered *closed-set face identification*, otherwise it is considered *open-set face identification*.

Closed-set face identification tasks are common in academic benchmarks, as galleries are carefully constructed by their authors to contain all probes. However, such problems are often not reflective of the real-world use of face recognition. In open-set face identification, a confidence threshold must be set to reject matches that do not meet a certain level of similarity. The selection of an appropriate threshold is especially relevant in high-risk applications such as law enforcement in which false positives have significant implications. In other cases, open-set face identification systems are used for *lead generation*, whereby a specific number of identities are returned, regardless of similarity, for human operators to manually review.

Face identification performance is reported in terms of accuracy in returning the correct identity of a probe from the gallery, or in the open-set case, no identity if the probe does not exist in the galley. Common performance metrics in closed-set face identification include *Rank-1 accuracy*, of all identification attempts, the number of times the correct identity in the gallery is the most similar identity to the probe and *Rank-10 accuracy*, the number of times the correct identity is in the ten most similar identities to the probe.

3.2.2 Deep Face Recognition

Rapid improvements in image classification in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [117] by AlexNet [72], ZFNet [148], GoogLeNet [127], and ResNet [54] from 2012 to 2015 cemented the DCNN as the standard method in computer vision research and applications. While early uses of convolutional neural networks in face verification showed preliminary success [16, 59], it was not until the introduction of the aforementioned network architectures that the modern era of deep face recognition began. Coupled with innovations in loss function design and access to larger image datasets, modern face recognition systems have improved state-of-the-art performance on benchmark face verification and identification tasks significantly in the past six years. For a complete survey of the development of deep face recognition systems, please refer to the review paper by Wang and Deng [134]; the following is a brief summary of major milestones.

The first system to adapt findings from ILSVRC to face recognition was Facebook's DeepFace [128], published in 2014 by Taigman et al. The nine-layer AlexNet-based model was trained on a private dataset of 4.4M images of 4K identities and achieved state-of-the-art accuracy on face verification tasks LFW and YouTube Faces (YTF) [139], reducing the error rate by more than 50% on the latter task.

Following this work, Google introduced FaceNet in 2015 with a major innovation in loss function design [118]. While the standard *softmax* loss function optimized inter-class differences, researchers found that intra-class differences remained high, problematic in the domain of face recognition. To rectify this problem, the *triplet loss* was introduced to jointly minimize the Euclidean distance between an anchor example and a positive example of the same identity and maximize the distance between an anchor and negative example. Using a ZFNet-based model and a private dataset of 200M images of 8M identities, they achieved state-of-the-art performance on LFW and YTF.

Innovations in loss functions dominated the next wave of improvements in benchmark tasks, motivated by improving discrimination between classes by making features more separable. Wen et al. introduced the Center Loss in 2016 [136], followed by Liu et al. with

the Angular Softmax in 2017 [82]. The Large Margin Cosine Loss was introduced in 2018 by Wang et al. [133], and in 2019, Deng et al. incorporated the Additive Angular Margin Loss into the ArcFace model [29], considered state-of-the-art on multiple face recognition benchmarks when published.

3.2.3 Face Recognition Training Datasets

Access to large-scale face recognition training datasets has been essential to the development of modern solutions by the academic community. While early works in the DCNN-era of face recognition came out of companies with access to massive private datasets, such as Facebook's 500M images and 10M identities [129] and Google's 200M images and 8M identities [118], the release of several open-source datasets in the ensuing years has allowed researchers to train models at scale. A summary of notable face recognition training datasets of the past six years is provided in Table 3.1. These datasets catalyzed the field of face recognition and lead to great advances in model performance on benchmark tasks. They largely consist of celebrity identities and copyrighted images scraped from the internet.

One exception is MegaFace, which is derived from the YFCC100M dataset of 100M photos with a Creative Commons Licence, from 550K personal Flickr accounts [131]. While the Creative Commons Licence permits the fair use of images, including in this context, Ryan Merkley, CEO of Creative Commons, noted the trouble of conflating copyright with privacy in a March 2019 statement,

> Copyright is not a good tool to protect individual privacy, to address research ethics in AI development, or to regulate the use of surveillance tools employed online. Those issues rightly belong in the public policy space, and good solutions will consider both the law and the community norms of CC licenses and content shared online in general. [90]

While MegaFace contains unknown, non-celebrity identities, an October 2019 investigation by the New York Times demonstrated that account metadata associated with images in the dataset allows for a trivial real-world identification of individuals [56]. In all datasets, no informed consent was sought or obtained for individuals contained therein.

Dataset	Year	Identities	Images	Consent Obtained	Source
CASIA WebFace	2014	10,575	494K	No	[147]
CelebA	2015	10,177	203K	No	[83]
VGGFace	2015	2,622	2.6M	No	[105]
MS-Celeb-1M	2016	99,952	10.0M	No	[48]
UMDFaces	2016	8,277	368K	No	[4]
MegaFace (Challenge 2)	2016	672,057	4.7M	No	[99]
VGGFace2	2018	9,131	3.3M	No	[15]

Table 3.1: Prominent open-source face recognition training datasets.

3.3 Ethical Considerations

3.3.1 Intent

The intent of this work is to investigate the performance of face recognition systems with respect to inclusion in training datasets. While one interpretation of this work may be to motivate efforts to mitigate demographic bias in the development of face recognition systems, it should be noted that increasing the performance of face recognition systems in any context can increase their ability to be used for oppressive purposes. In addition, due to historical societal injustices against marginalized populations and racially-biased police practices in the United States, a disproportionate number of Black and Hispanic people are present in mugshot databases, often used by law enforcement agencies as data sources for face recognition systems [98, 41]. These populations are therefore poised to receive a greater burden of the effects of improved face recognition systems. I therefore position this work as informing the discussion on data privacy and consent when it comes to face recognition systems and do not advocate for technical improvements without a larger discussion on the appropriate use and legality of the technology.

3.3.2 Use of MS-Celeb-1M

As previously noted, the MS-Celeb-1M dataset was removed from Microsoft's website in June 2019. In a response to a Financial Times inquiry, Microsoft stated the website was retired "because the research challenge is over" [97]. However, a version of this dataset with detected and aligned faces from a "cleaned" subset of the original images is available from the Intelligent Behaviour and Understanding Group (iBUG) at Imperial College London.

The dataset was offered as training data for the "Lightweight Face Recognition Challenge & Workshop"[1] the group organized at ICCV 2019. The group has pre-trained face recognition models available as benchmarks for the challenge, trained on this data.

As this work aims to conduct experiments in a realistic setting in order to better inform the conversation around data collection processes, the use of a state-of-the-art model trained on a large dataset is necessary to gain insights that are applicable to commercial applications. I therefore use the MS-Celeb-1M dataset, through its derived version offered for the ICCV 2019 Workshop, for the limited scope of this work.

3.4 Methodology

3.4.1 Face Recognition Model

Training Data

A cleaned version of the MS-Celeb-1M dataset [48] is used as training data for a face recognition model in this work. This dataset was prepared for the ICCV 2019 Lightweight Face Recognition Challenge [30]. All face images were preprocessed by the RetinaFace model for face detection and alignment [28]. A similarity transformation was applied to each detected face using five predicted face landmarks to generate normalized face crops of 112 x 112 pixels.

As the original version of this dataset has been shown to exhibit considerable inter-class noise, efforts have been made to automatically clean the dataset [63]. In the case of this version, after face detection and alignment, cleaning was performed by a semi-automatic refinement strategy. First, a pre-trained ArcFace model [29] was used to automatically remove outlier images of each identity. A manual removal of incorrectly labelled images by "ethnicity-specific annotators" followed to result in a dataset of 5,179,510 images of 93,431 identities. This dataset is referred to as *MS1M-RetinaFace*.

Model

The ArcFace model [29] is used in this work. ArcFace employs the Additive Angular Margin Loss and a ResNet100 backbone to arrive at a 512-dimensional feature representation of

[1] https://ibug.doc.ic.ac.uk/resources/lightweight-face-recognition-challenge-workshop/

an input image. The model achieves a verification accuracy of 99.83% on LFW and Rank-1 identification accuracy of 81.91% on the MegaFace Challenge 1 with 1M distractors, considered state-of-the-art results. This model is selected for study as is the top academic, open-source entrant on the National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) 1:N Verification[2], comparable with many commercial vendors' systems. NIST's FRVT is the gold standard benchmark in testing face recognition systems in 1:1 and 1:N searches. Pre-trained weights for this model were provided by iBUG.

3.4.2 Experiments

To determine the effect of inclusion in the training data of a face recognition system on its ability to identify an individual, the problem is framed as a closed-set face identification task. Two probe datasets are constructed and face identification experiments are performed on a gallery of 1M distractor images. Performance of the model on the probe datasets is evaluated in terms of Rank-1, Rank-10, and Rank-100 identification accuracies. A visual representation of the datasets used in this work is shown in Figure 3.1.

Probe Data

Two probe datasets are constructed from the VGGFace2 dataset [15]. Using regular expressions, identities in VGGFace2 are matched by name with the identify list of MS1M-RetinaFace. 5,902 VGGFace2 identities are found to be present in MS1M-RetinaFace and 3,229 VGGFace2 identities are not present in the training dataset. In each of these two groups, 1,000 male identities and 1,000 female identities are randomly selected for evaluation, based on gender labels provided by VGGFace2 metadata. For each identity, 50 images are randomly selected and undergo face detection and alignment by the Multi-task Cascaded Convolutional Network (MTCNN) [149] to generate normalized face crops of 112 x 112 pixels. The set of 100,000 images of 2,000 identities present in the training data is referred to as the *in-domain probe set* and the set of 100,000 images of 2,000 identities not present in the training data is referred to as the *out-of-domain probe set*. Finally, 512-dimensional feature representations for all images in the in-domain and out-of-domain probe sets are generated by the ArcFace model.

[2] https://pages.nist.gov/frvt/html/frvt1N.html

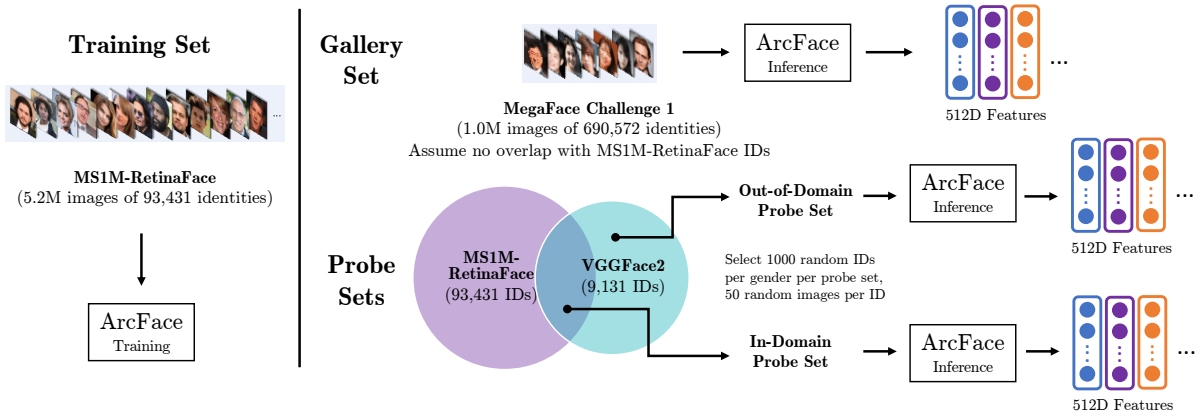

Figure 3.1: Experimental procedure to generate feature representations of images in gallery and probe sets from ArcFace model.

Gallery Data

The MegaFace Challenge 1 Distractor dataset [69] of 1,027,058 images of 690,572 identities is used to form the basis of the *gallery*. MTCNN is used to generate normalized face crops of 112 x 112 pixels and ArcFace is used to generate 512D feature representations of each image in the gallery.

Evaluation Protocol

The experiments conducted in this work follow the protocol of MegaFace Challenge 1, with the two novel probe sets in place of the standard FaceScrub test set [101]. The Linux development kit offered by MegaFace is used to perform evaluation. Each probe set is evaluated following Algorithm 1; a written description of this protocol follows.

A probe set contains 2,000 identities, each with 50 images represented as 512D features. For each identity, we iterate over their 50 images, adding one image to the gallery at a time, which is referred to as *the needle*. We then iterate over the remaining 49 images, using each one as a probe. All images in the gallery are ranked by Euclidean distance in feature space to the probe. An individual's Rank-1, Rank-10, and Rank-100 face identification accuracy is the number of times the needle is within the top 1, 10, and 100 positions in the ranked list, respectively, across their 2,450 searches (50 needles $\times$ 49 probes per needle). Results

Algorithm 1: Closed-set face identification evaluation.

```
Result: Rank-1, 10 and 100 face identification accuracies for a probe set.
Rank-1_Acc., Rank-10_Acc., Rank-100_Acc. = empty lists;
gallery contains 1M distractor images;
for identity in identities_1 to 2000 do
    r_1, r_10, r_100 = 0;
    for image_needle in images_1 to 50 do
        add image_needle to the gallery;
        for image_probe in images_1 to 50 do
            if image_needle == image_probe then
                continue;
            else
                rank all images in gallery by Euclidean distance to image_probe in
                 feature space;
                if image_needle in first position in ranked list then
                    r_1 = r_1 + 1
                if image_needle in first 10 positions in ranked list then
                    r_10 = r_10 + 1
                if image_needle in first 100 positions in ranked list then
                    r_100 = r_100 + 1
        remove image_needle from gallery;
    r_1 = r_1/(50 × 49); add r_1 to Rank-1_Acc.;
    r_10 = r_10/(50 × 49); add r_10 to Rank-10_Acc.;
    r_100 = r_100/(50 × 49); add r_100 to Rank-100_Acc.;
Rank-1_Acc. = average of Rank-1_Acc.;
Rank-10_Acc. = average of Rank-10_Acc.;
Rank-100_Acc. = average of Rank-100_Acc.;
```

for a probe set are reported as the average Rank-1, Rank-10, and Rank-100 accuracies across their 2,000 identities.

To determine if the face identification accuracies of two probe sets at a certain rank differ significantly, a two-sample, two-sided independent Welch's t-test is employed [135]. If the two samples are drawn from distributions with different means with a strict threshold of a p-value less than 0.01, the probe sets are said to exhibit a significant difference in performance.

3.5 Results and Discussion

Results of the experiments are presented in Table 3.2 for Ranks 1, 10, and 100. There appears to be an increase in face identification accuracy for identities present in the training data, compared to those who are not. In-domain identities have a 4.0% higher identification accuracy than out-of-domain identities at Rank-1 (79.7% vs. 75.7%), and are 4.3% higher at Rank-10 (91.0% vs. 86.7%) and 3.6% higher at Rank-100 (92.9% vs. 89.3%), all of which are significant findings. These results suggest that modern DCNN-based face recognition systems are biased towards individuals they are trained on.

The disparate performance between probe sets suggests some amount of overfitting has occurred in the model. Although the model generalizes well to new identities, as evidenced by results on benchmarks LFW, MegaFace and on NIST's FRVT, these results indicate that the 93k identities the system is trained on are more easily identifiable in a large-scale study. As the model's Additive Angular Margin Loss sought to increase discrimination between classes by making features more separable, it appears the model has learned to map identities to the same feature representation more consistently for those it has seen before.

The role of gender in the performance of the face recognition model is also investigated. Within in-domain identities, no significant difference between males and females is reported. However, within out-of-domain identities, a significant decrease in performance is reported for females compared to males across all ranks. These results suggest that the model does not generalize well to new female identities. Indeed, the largest drop in performance between probe sets is exhibited between domains for female identities, as significant decreases of 6.7%, 6.2%, and 5.2% are reported at Rank-1 (80.9% vs. 74.2%), Rank-10 (90.8% vs. 84.6%), and Rank-100 (93.1% vs. 87.9%), respectively. While gender labels are not available for all identities in MS1M-RetinaFace, recent work has demonstrated that large-scale face recognition datasets greatly over-represent lighter-skinned males [91]. A representational bias in MS1M-RetinaFace may account for the model's disparate ability to generalize to new female identities. Looking at these results in a different way, the relatively consistent performance between males and females in the in-domain probe set is perhaps more evidence that the model is overfitting to identities it has seen before. If the model only exhibited a gender bias against females, we would expect to see this in both in-domain and out-of-domain probe sets. However, in-domain results remain consistent between males and females, yet the model struggles most to identify novel female identities. These results suggest that the model exhibits a "training inclusion bias" that is more pronounced than its gender bias.

Results of this study lead to the question: is the bias towards individuals in training

Probe Set	Rank-1 Acc.	Rank-10 Acc.	Rank-100 Acc.
In-Domain	0.797*	0.910*	0.929*
Out-of-Domain	0.757*	0.867*	0.893*
In-Domain Males	0.785	0.911*	0.927*
Out-of-Domain Males	0.773	0.888*	0.907*
In-Domain Females	0.809*	0.908*	0.931*
Out-of-Domain Females	0.742*	0.846*	0.879*
In-Domain Males	0.785	0.911	0.927
In-Domain Females	0.809	0.908	0.931
Out-of-Domain Males	0.773*	0.888*	0.907*
Out-of-Domain Females	0.742*	0.846*	0.879*

Table 3.2: Face identification accuracies of ArcFace model on different probe image sets with 1M distractor images. (*) denotes significance between probe sets at $p < 0.01$.

data truly a consequence of overtraining, or is this a fundamental element of deep face recognition models? Overfitting in a traditional sense seems unlikely, as early stopping was employed during the training phase, and results on held-out test identities demonstrate strong generalization. Perhaps there is a generalization gap in performance between in-domain and out-of-domain identities that is not observable given current evaluation methods, and increased regularization can mitigate this gap. Further testing on different training datasets and model architectures will be necessary to gather more evidence to answer this question.

The effect of Fitzpatrick skin type [40] on face recognition model performance was not evaluated in this study, as skin type annotations were not available. However, two considerations were made to attempt to control for effects of skin type in these results. First, the selection of 2,000 identities for each probe set is far larger than what is used in the standard protocol of MegaFace Challenge 1, where 80 identities are sampled from FaceScrub [69]. Having a larger sample size helps to control for identities who may have either superior or poor performance due to possible model bias. In addition, the approach of random sampling in-domain and out-of-domain probe sets ensures that both contain a similar distribution of identities with respect to skin type, with the assumption that the identities common to MS1M-RetinaFace and VGGFace2 and the identities distinct to VGGFace2 follow the same distribution of skin type. As both MS1M-RetinaFace and VGGFace2 use the popularity of celebrities online to construct identity lists, this assumption seems

reasonable. Having said this, the role of skin type in the performance of the model is a very important relationship to study in future work. Fitzpatrick skin type annotations will need to be collected for all individuals in VGGFace2 such that sampling can be done to ensure even representation in probe sets across gender and skin type and to determine intersectional accuracy.

The results of this study are quite concerning from a privacy and informed consent perspective. As described in 3.2.3, there does not exist a major open-source dataset that gathers informed consent from the individuals it contains. Without these individuals' knowledge or permission, the systems trained on their identities have a greater ability to identify them. As face recognition becomes more powerful and ubiquitous, the ability for misuse becomes greater. While MS-Celeb-1M contains only "celebrity" identities, this classification of an individual should not negate informed consent in the development of powerful surveillance technologies. Face recognition systems are distinct among biometrics as the face uniquely identifies a person with high discriminability, yet a high-quality image of the face can easily be captured at a distance without one's knowledge, cooperation, or consent. It is difficult to opt-out of these systems without wearing a mask or other means of obfuscation, drawing undue attention to one's self.

3.6 Chapter Summary

The key findings from this chapter are:

- Face recognition systems have advanced rapidly over the past six years through the application of deep learning methods such as the CNN to the problem space.
- CNNs require large amounts of data to accurately train a model that can produce feature representations with small intra-class and large inter-class variability.
- Open-source face recognition training datasets in the academic community have similarity increased in scale during this period, by virtue of methods that collect millions of images of hundreds of thousands of individuals from the web.
- Informed consent in the collection of individuals' face images, however, has been completely absent from these efforts.
- Through a novel face identification experiment modelled after the MegaFace Challenge 1, a state-of-the-art face recognition model was found to exhibit a significantly

greater ability to identify an individual in a 1:N search if their identity was included in the training data of the model ($p < 0.01$).

While the notion of a person having their identity included in a face recognition training dataset without their consent seems uncomfortable in an abstract manner, this work is the first to suggest tangible evidence of a disparate performance impact for individuals in 1:N searches by models trained on their identity. Although this work is limited by the availability of open-source training datasets and face recognition models, its methodology was designed to simulate a real-world testing environment of a state-of-the-art face recognition system, with a gallery of more than 1M images and an experimental procedure that saw 9.8M 1:N searches conducted. These findings, therefore, may hold for systems currently deployed in the world. As there exist hundreds of vendors who manufacture and sell face recognition systems to an open market[3], it is concerning that there is no requirement for the disclosure of the source and contents of the training data that underpins their technologies. As face recognition systems represent a contentious dual-use technology that can enable mass surveillance, this work hopes to better inform the research community of the consequences of ignoring consent in web-scraped data collection, and to offer evidence that transparency into the training data of vendors of such systems is needed.

[3] https://pages.nist.gov/frvt/html/frvt1N.html

Chapter 4

Label Taxonomy in Facial Expression Recognition: Re-contextualizing the Problem

4.1 A Canonical Dataset, Revisited

The Japanese Female Facial Expression (JAFFE) dataset [84] is a canonical dataset in facial expression recognition research and development. Collected in 1998 by Michael Lyons, Miyuki Kamachi and Jiro Gyobaat in the Psychology Department of Kyushu University, the dataset is comprised of 213 images of 10 Japanese females posing in "the six basic facial expressions" of "happiness", "sadness", "surprise", "anger", "disgust", and "fear", as well as in a "neutral face." The photos were taken in a controlled lab environment by the expressors themselves as they peered through a semi-reflective plastic sheet towards the camera. For each image, 60 female Japanese undergraduate students scored the degree to which each of the six basic facial expressions were present on a 5-point scale. Representative images from JAFFE are displayed in Figure 4.1.

JAFFE was perhaps the first dataset freely circulated online for non-commercial research in facial expression recognition. It has been used in more than 2,000 publications and remains relevant today with the advent of CNN-based classification systems [77]. By providing a standardized dataset, JAFFE helped to formalize the evaluation of new algorithms in facial expression recognition in seven posed classes. A closer look at JAFFE, however, uncovers a dataset that is fraught with issues in data collection, annotation, and use that beg broader questions about the nature of the problem it is addressing.

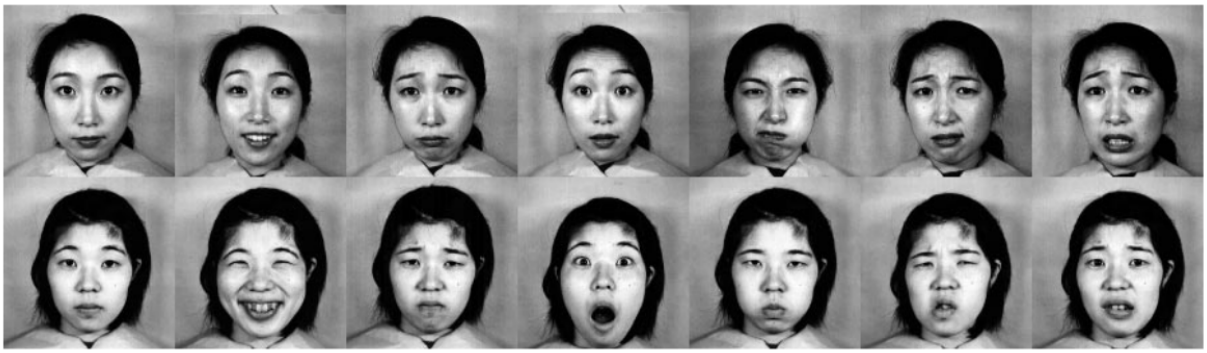

Figure 4.1: A sample of images from the JAFFE dataset. From left to right; neutral, happiness, sadness, surprise, anger, disgust, fear [85].

With this context, this chapter aims to address the research question: how do canonical facial expression recognition datasets hold up to new research into the nature of emotion and facial expressions, namely, Barrett et al.'s [7] criteria for identifying emotional inference from facial expression images?

4.1.1 Into the Label-Verse

The common convention in using JAFFE in research is to classify each face into one of the seven posed classes (six basic facial expressions plus neutral). Of the 50 research papers identified to have used JAFFE in facial expression recognition from January 1 to June 30, 2020, 46 employed this technique. The remainder, however, used the dataset in a different manner. These works are listed in Appendices A and B, respectively. As previously noted, each image in JAFFE was also labelled by a large cohort of annotators who provided scores for the extent to which each of the six emotions were present in the face. This semantically-annotated version JAFFE, or Scored JAFFE (S-JAFFE) as it has become known, was seldom used in the years after its publication, but the introduction of "label distribution learning" (LDL) by Geng et al. in 2013 set the stage for its re-introduction [45]. LDL is a paradigm in which the degree to which a certain set of labels representing an instance is learned as a distribution. This paradigm was taken up by Zhou et al. in 2015 who applied it to facial expression recognition as "emotion distribution learning," with S-JAFFE being one of two datasets used in experiments [153].

The scores for each of the six annotated facial expression classes in S-JAFFE are pro-

vided as averages in the dataset; the individual annotators' assessments no longer available as they have been lost to time.[1] Although inter-rater reliability can not be analyzed through metrics such as Cohen's kappa coefficient [18], the provided averages are still enlightening. One would expect these annotations to align well with the posed expressions, however, this is not always the case. Figure 4.2 presents an image of the posed facial expression "disgust" with its semantic scores and normalized semantic scores, computed by scaling each raw score to be between 0 and 1 and normalizing the score vector to equal 1. It can be observed that for this instance, the annotators saw it as presenting a stronger expression of "anger" than the posed expression of "disgust." Indeed, of the 183 images that were posed in one of the six basic facial expressions, 35 (19%) were observed to express a different class as the strongest. Further, by averaging the normalized semantic scores across all examples in a class, a sense of the annotators' overall view of the posed facial expressions can be observed. The results, displayed in Table 4.1, demonstrate that annotators did not see each class as a distinct expression of the posed emotion, but as a composite of many. While the posed class of "happiness" is relatively strongly observed (65.63%), the posed class of "fear" is observed as almost equally a composite of "sadness", "surprise", "disgust" and "fear" (~20% each). In the context of these results, the conventional use of JAFFE in a single-label classification problem seems to conflict with observations of the annotators. Multiple surveys in the domain of deep learning-based facial expression recognition fail to acknowledge this inconsistency [77, 152].

It should be noted that the authors of JAFFE did address the issue of inconsistent labels, albeit in a troubling manner. In their paper introducing the dataset, Lyons et al. [85] proposed a novel method to code facial expressions with Gabor filters. Results of their method on posed faces with the "fear" expression, however, were poor. In response, they simply decided to remove these images from their experiments, employing a new cohort of annotators to re-score each image with the remaining five facial expression classes and recomputing results. On this topic, the authors wrote in their paper, "Model/Data agreement was higher with fear excluded. Fear is considered to be a problematic expression for Japanese expressors and subjects for reasons beyond the scope of the present article" [84]. And again, in a document provided with the dataset, they wrote,

> This rating experiment excluded fear images and and the fear adjective from the ratings. We did this because we thought that the expressors were not good at posing fear. There is some evidence in the scientific literature that fear may be processed differently from the other basic facial expressions. [84]

[1] Personal correspondence with Michael Lyons, June 8, 2020.

Expression	Raw Score	Normalized Score (%)
happiness	1.43	4.15
sadness	2.87	18.03
surprise	1.77	7.43
anger	4.33	32.11
disgust	3.87	27.68
fear	2.10	10.61

Figure 4.2: Raw and normalized semantic scores for an example image in the posed facial expression "disgust".

Posed Class	% happiness	% sadness	% surprise	% anger	% disgust	% fear
neutral	25.54	19.30	13.27	15.33	13.94	12.61
happiness	65.63	9.00	9.57	4.94	5.16	5.70
sadness	5.88	33.06	7.24	14.4	21.20	18.13
surprise	15.87	10.16	41.55	9.08	9.35	13.98
anger	5.59	15.87	7.80	33.45	26.90	10.39
disgust	4.48	16.23	9.87	21.05	32.63	15.74
fear	4.93	18.31	22.31	12.85	20.23	21.36

Table 4.1: Average normalized semantic scores for posed facial expression classes in JAFFE.

No references were provided to substantiate the claim of "fear" being more difficult to produce, identify, or process. In addition, images from one expressor under-performed on the authors' proposed method compared to the other expressors, so they decided to also remove these images from their results, stating, "Expressor NM was considered to be an outlier and excluded from the above quoted averages and ranges. On inspection NM's expressions were difficult to interpret" [84]. Should the images of an individual's expressions that deviate from normative expectations be considered as outliers, or be referred to as "erratic," as the authors do later in their paper? Perhaps instead of changing the data used to validate their facial expression coding method, the framing of the problem needs to be altered.

4.1.2 The Implicit Assumptions of JAFFE

As discussed by Crawford and Paglen in their essay *Excavating AI* [23], there exist many implicit assumptions embedded in the label taxonomy of the JAFFE dataset, and in the problem of facial expression recognition at large:

- There exist six basic emotions.
- There is a relationship between one's facial expression and their inner emotional state.
- Across emotional episodes, an individual will produce the same facial expression.
- Across individuals, the same facial expression will be produced for a given emotional state.
- Observers can readily interpret a person's facial expressions to assess their inner emotional state.
- A single image of the face is sufficient to assess one's inner emotional state.

Lyons et al. cite a 1975 book by Paul Ekman and Wallace V. Friesen [35] to justify the chosen label taxonomy of JAFFE, and to a large extent, the assumptions listed above. Indeed, work by Ekman in the second half of the 20th century popularized the view of emotions as six discrete entities, the same across individuals and cultures, and clearly discernable from the face [34]. This view of emotion, referred to as the "common view" [7], is widely employed in computer vision research and embedded in facial expression recognition datasets collected in laboratory conditions (CK+, MMI, Oulu-CASIA) and scraped from the web and annotated as "in-the-wild" collections (FER-2013, RAF-DB, AffectNet) [77]. Microsoft, Google, and Amazon all offer facial expression recognition APIs that embed the assumption of discrete, universal, and discernable emotions into

their products. Despite the wide adoption of this line of research, it has been deeply contested over the years by psychologists, anthropologists, and other researchers [137]. In July of 2019, a major review study lead by Lisa Feldman Barrett and an interdisciplinary group of experts in psychology, social sciences, and computer science was published in *Psychological Science in the Public Interest*, making a strong case that the field of facial expression recognition, as it currently exists, sits on a faulty scientific foundation [7].

4.2 Facial Expressions, Reconsidered

In their review, Barlett et al. [7] analyzed over 1,000 studies across psychology, neuroscience, computer vision, and other domains that studied emotion expression, focusing on evidence pertaining to the six emotion categories of "happiness", "sadness", "surprise", "anger", "disgust", and "fear". They summarized evidence in two discrete areas: how people actually move their face during emotional episodes, *expression production*, and which emotions are actually inferred from looking at facial movements, *expression perception*. The researchers assessed how each of these areas held up against the criteria of *reliability*, *specificity*, *generalizability*, and *validity*.

Reliability measures how well instances of the same emotion category are expressed or perceived from a common set of facial movements. When a person is "sad", for example, how often is a specific expression, such as a scowl, produced? Or, when a person makes a frowning facial configuration, for example, how often are they perceived to be "sad"? Specificity measures how uniquely a facial configuration maps to an emotion, or how uniquely perceivers rate a certain facial expression as a corresponding emotion. To be considered the expression of "anger", a scowling face must not be similar to any other class of expressions. Or, if a frowning face is perceived as the expression of "sadness", then it should be labelled as uniquely "sad". Generalizability measure how well rates of reliability and specificity are replicated across studies in both production and perception, particularity across different populations of people. In their work, Barrett et al. assess evidence from healthy adults from the United States and other developed nations, healthy adults living in small-scale remote villages, healthy infants and children, and people who are congenially blind [7]. Validity measures the extent to which a person is *actually experiencing the expected emotional state*, preferably measured by an objective metric. Even if strong generalizability exists in production and perception, validity is necessary to confirm that an inference into one's inner emotional state is well-founded. In discussing these criteria, Barlett et al. posit,

> If any of these criteria are not met, then we should instead use neutral, descriptive terms to refer to a facial configuration without making unwarranted inferences, simply calling it a smile (rather than an expression of happiness), a frown (rather than an expression of sadness), a scowl (rather than an expression of anger), and so on. [7]

By almost all accounts, the researchers find that the "common view" of facial expressions does not hold up to their rigorous review of empirical evidence [7]. They find limited reliability, a lack of specificity, limited generalizability, and an overall dearth of research into validity in production and perception studies. Their findings suggest that people may smile when "happy" and frown when "sad", perhaps more than chance would expect, but the manner by which people communicate these emotions varies substantially across cultures, situations, and even by individuals in a single situation. While a scowling face may be *an* expression of "anger" in certain instances, it is not *the* expression of "anger" in any generalizable or universal way. Indeed, they state that "prototypical expressions" in the six basic emotions are best thought of as stereotypes, failing to capture the rich variety with which people spontaneously move their faces to express emotions in daily life. On the topic of facial expression recognition by computer vision practitioners, the authors offer the following analysis,

> [T]ech companies may well be asking a question that is fundamentally wrong. Efforts to simply 'read out' people's internal states from an analysis of their facial movements alone, without considering various aspects of context, are at best incomplete and at worst entirely lack validity, no matter how sophisticated the computational algorithm. [7]

Returning to JAFFE, it is apparent how this dataset fails to satisfy the criteria for emotional inference. The dataset lacks reliability, as evidenced by the expressor who is said to have produced "erratic" expressions. It lacks specificity, as demonstrated by the extent to which each posed class is perceived to encompasses multiple facial expressions, most dramatically with the "fear" class. Generalizability is not well considered, as the dataset is of a very specific population of individuals, namely young Japanese females. And lastly, the validity of the facial expressions is not investigated, although the fact that they were posed, without an external stimulus or an attempt to solicit a spontaneous emotional episode, appears to be evidence enough to fail this criterion.

4.3 Label Taxonomy, Reassessed?

4.3.1 Slow Progress

The response to this landmark review in the computer vision community has been mixed. At CVPR 2020, the leading conference in computer vision research, a workshop entitled "Challenges and Promises of Inferring Emotion from Images and Video" was organized by Aleix M. Martinez, a co-author of the study. The workshop invited Barrett to give a talk on the topic "Can Machines Perceive Emotion?" Organizers also included the following statement in their Call for Papers,

> Recent research shows that faces or body expressions alone are insufficient to perform a reverse inference of image to emotion, and that context, personal believes, and [culture] must be accounted for. This workshop will present these limitations and examine several alternative approaches to successfully interpret the emotion and intent of others.[2]

However, upon review of the works published at this workshop, there appears to be limited scholarship on the aforementioned challenges, as most works continue to engage with only images of faces in their study of emotion.

On the data front, there are recent efforts to include more contextual signals in emotion recognition datasets. The EMOTions In Context (EMOTIC) dataset, for example, contains images with people in real environments, annotated with their apparent emotions [74]. These datasets, however, still suffer from many of the same issues that come with annotating the perceived emotional state of others in images. Despite efforts to shift focus in computer vision away from the use of static images of the face in recognizing emotion, new methods continue to be developed in emotion recognition that rely on antiquated datasets [46, 81].

Even when the issues previously discussed are understood by practitioners, there remains resistance to change course from the "common view." On August 22 2019, one month after the publication of the review by Barrett et al., Amazon quietly updated the documentation for their Emotion API. They changed the description of the API's return value from "The emotions detected on the face" to "The emotions that appear to be expressed on the face", adding, "The API is only making a determination of the physical appearance of a person's face. It is not a determination of the person's internal emotional state

[2] http://cbcsl.ece.ohio-state.edu/cvpr-2020/index.html

and should not be used in such a way."[3] While this description acknowledges that their solution is laced with issues, it does not change the mechanics of their product. Indeed, their API continues to interpret a face and return scores for the presence of the classes `HAPPY, SAD, ANGRY, CONFUSED, DISGUSTED, SURPRISED, CALM, UNKNOWN,` and `FEAR,` so it is doubtful their disclaimer had any meaningful impact on the use of their product.

There exists an interesting possibility that a new taxonomy has not been widely adopted in facial expression recognition due to limitations of the English language. There simply does not exist concise terminology to describe the stereotypical facial expressions of "surprise," "disgust," and "fear" in the same way that "a smile" does for "happiness," "a frown" does for "sadness," and "a scowl" does for "anger." This lack of vocabulary encourages the continued use of emotional language to describe faces, which brings with it a conflation of emotions and specific facial configurations.

4.3.2 Harms of Current Taxonomy

Language is important. The taxonomy we use in datasets frames the problems we solve. We should always strive to tackle problems that have a strong foundation in science. And as the evidence has shown, the classification of face images labelled with emotions does not. Although models may achieve good performance on these datasets, it does not mean they are accurate. A facial expression recognition model does not learn what "happiness" is in a face, it learns to detect the common facial configuration of images labeled "happiness," which is often a smile. We should be specific with taxonomy to ensure the problems we are addressing are accurately described. Re-labelling existing datasets with better annotations such as facial landmarks and facial muscle activity, and resisting the urge to map these annotations back to an emotional state, is one approach for moving forward.

Before proceeding, however, we need to reconcile with the present. The way facial expression recognition technology is being developed and deployed in society today is troubling. These systems lack the reliability and specificity to interpret facial expressions, yet this has not hindered their commercialization and application to sensitive areas of life. Firms HireVue and VCV sell facial expression recognition technology to assess candidates during video job interviews for "employability" through microexpressions and visual and auditory cues [53]. Technology developed by Hikvision has been used in Chinese middle schools to determine the emotional state and engagement level of students, with each pupil receiving a real-time "attentiveness" score during class [76]. In the UK, startup WeSee is working with law enforcement to analyze suspects in interviews with officers. In 2018, their

[3] GitHub commit history for Amazon Emotion AI.

CEO David Fulton told the BBC, "Using only low-quality video footage, our technology has the ability to determine an individual's state of mind or intent through their facial expressions, posture, gestures and movement" [130].

In hiring, education, and law enforcement, a technology that is not founded in science is assisting in decision-making that materially impacts people's lives. To make matters worse, issues of gender and racial bias may exacerbate the already fraught situation. As demonstrated by Buolamwini, Gebru, and Raji in *Gender Shades* and its follow-up work *Actionable Auditing* [49, 108], darker-skinned people, especially darker-skinned women, are often subject to much worse performance from facial analysis technologies than their light-skinned peers. Indeed, a 2018 study by Rhue found racial biases in facial expression recognition services provided by Face++ and Microsoft, with both systems interpreting Black NBA players to have more negative emotional states than their white colleagues on images controlled for facial expressions [112]. Without mechanisms to audit facial expression recognition systems, historically marginalized groups may see a greater burden of this harmful technology.

4.4 Chapter Summary

The key findings from this chapter are:

- The Japanese Female Facial Expression (JAFFE) dataset was one of the first widely-available datasets for non-commercial research in facial expression recognition, and by virtue of this distinction, it helped to formalize the evaluation of facial expression recognition algorithms on images of individuals posing in "the six basic emotions" of "happiness", "sadness", "surprise", "anger", "disgust", and "fear".
- A close examination of the paper introducing JAFFE, however, demonstrates that (i) emotions perceived by annotators did not reliably match the posed expressions of expressors, and (ii) authors of JAFFE took many liberties in reporting results of a novel algorithm on the dataset by removing expressors and emotional categories that were too unstable for their liking, suggesting there exist fundamental issues with the integrity of the data.
- JAFFE and much of the field of facial expression recognition sits atop a body of research that asserts emotions are discrete, enumerable entities that manifest in the face in the same manner across individuals and cultures, however, a comprehensive

2019 review of over 1,000 studies in psychology, neuroscience, computer vision, and other domains strongly refutes these assertions.

- Viewing JAFFE through the lens of this new work brings to light issues in the reliability, specificity, generalizability, and validity of the dataset, leading to the idea that its label taxonomy should be changed from one that that makes reverse inferences of emotion to one that uses neutral, descriptive terms to refer to the facial configurations of expressors.

- While some discourse has begun in the computer vision community to move away from static face images in emotion research, response to this landmark study has largely been muted as many works continue to perpetuate the "common view" of emotion as a valid scientific stance, bolstering harmful applications of facial expression recognition in applications that have meaningful impacts on individuals.

This work aims to motivate computer vision practitioners to reconsider the label taxonomy of their datasets, as uninformed research can support the commercialization of unscientific technology. Although facial expression recognition systems have begun to be deployed, the widespread use of this technology does not have to be inevitable. In their 2019 Annual Report, The AI Now Institute at New York University made their first recommendation, "Regulators should ban the use of affect recognition in important decisions that impact people's lives and access to opportunities. Until then, AI companies should stop deploying it" [22]. In January 2020, The Artificial Intelligence Profiling Act (House Bill 2644) was introduced in the Washington State Legislature. The bill prohibits the operation or installation of equipment that incorporates AI-enabled profiling in public places, and prohibits the use of such technology to make decisions that produce legal effects or similarly significant effects in criminal justice and employment, among other areas [121].

However, without regulatory measures, it is difficult to imagine the proliferation of this technology slowing down. Some, however, have provided an idea of what its widespread deployment may look like, in an effort to guide us to actions that can mitigate this future,

> It was terribly dangerous to let your thoughts wander when you were in any public place or withing range of a telescreen. The smallest thing could give you away. A nervous tick, an unconscious look of anxiety, a habit of muttering to yourself-anything that carried with it the suggestion of abnormality, of having something to hide. In any case, to wear an improper expression on your face (to look incredulous when a victory was announced, for example) was itself a punishable offence. There was even a word for it in Newspeak: FACECRIME, it was called. — George Orwell, *1984* [103]

Chapter 5

Discussion and Conclusions

Undesirable biases, non-consensual data collection, and inappropriate label taxonomies are rife in computer vision datasets. Through a multi-pronged approach, this thesis has critiqued normative practices in data collection and use, emphasizing harms that come with the abstraction of data.

ImageNet's impact on the field of computer vision is undeniable, but the practices it endorses are untenable. We see elements of each of the three themes of this thesis embedded in ImageNet. Representational bias with respect to gender and age were identified in the dataset in Chapter 2, as only 41.62% of people in ILSVRC-2012 present as feminine, 1.71% appear to be over the age of 60, and the largest group represented is masculine presenting individuals aged 15 to 29 at 27.11%. This masculine skew manifests in classes related to academia, business, and sports, promoting a male-centred view of these disciplines. Consent was not obtained from the human subjects in ImageNet during collection, which is particularity concerning in the case non-consensual photographs included the `miniskirt` class, as identified by Prabhu and Birhane [107]. The taxonomy of labels in ImageNet is a major problem, particularity in the `person` subtree. Some classes, such as `good person`, are not visually grounded and lead to the perpetuation of stereotypes and physiognomy, while others, such as `nymphet`, defined as "a sexually attractive young woman," are overtly offensive, using sexist language that promotes the objectification of women. Although the issue of copyright in datasets is largely out of scope of this thesis, it remains an additional contentious issue in modern data collection. While the terms of ImageNet call for its use only in "non-commercial research and educational purposes,"[1] its widespread use by computer vision startups and large companies has rendered this restriction, and

[1] http://image-net.org/download-faq

any underlying protections provided by fair use doctrine, moot. ImageNet is emblematic of a community that prioritizes getting algorithms to work over ethical considerations of their work [5], and one that often only becomes aware of issues of bias from scholars from underrepresented populations in the community [13, 108, 119, 114, 124].

Face recognition and facial expression recognition are two domains in computer vision that also intersect with the main themes of this thesis. In both cases, biases in datasets and derived models have been well-documented [49, 91, 112], leading to disproportionate performance impacts for women and non-white individuals. However, even when these systems mitigate bias and work as intended, they still pose risks for harm. Face recognition requires large "galleries" of images to search through, which are largely concentrated with those in positions of power. This technology is unregulated in Canada and most parts of the world [32], allowing law enforcement agencies to use it in their daily work without transparency or governance. Findings from Chapter 3 present evidence that state-of-the-art face recognition systems have disparate identification accuracies for individuals dependent on their inclusion in training data, which is collected without individuals' knowledge or consent. As there exists no requirement for prospective vendors to disclose the details of their training datasets when selling to law enforcement entities, these findings are troubling. With a documented history of systemic racism in the Canadian criminal justice system [95, 89], reports of the technology being used to identify protesters around the world [111, 12, 96], and the potential to enable mass surveillance, face recognition poses many risks for misuse. Facial expression recognition poses similar risks, although not specifically through implications of the non-consensual nature of data collection. As discussed in Chapter 4, the taxonomy of datasets in this domain are not well-founded in science. As these systems are increasingly being deployed to help make decisions in consequential areas of daily life [53, 76, 130], continued research into inferring one's emotional state from static images of their face bolsters the unproven claims that vendors advertise.

As this thesis demonstrates, computer vision practitioners largely do not stop to interrogate their assumptions in data collection and use. With increasingly large datasets being used in industry research, this trend is bound to continue if due diligence is not paid to data. Google researchers often employ an internal dataset called JFT-300M in their work, comprised of 300 million images sourced from the web [57, 125, 143], while Facebook researchers employ images from Instagram in the billions to train models [86].[2] In a blog post announcing a 2017 study [125], Google researchers wrote, "Furthermore, building a dataset of 300M images should not be a final goal - as a community, we should explore if models continue to improve in a meaningful way in the regime of even larger (1 billion+

[2]In one experiment, researchers trained a CNN on 3.5B images, using 336 GPUs for 22 days, costing approx. $129,000 USD [75] to train a single model.

image) datasets".[3] This statement is emblematic of the idea of surveillance capitalism, an economic system Zuboff describes the commodification of personal data with the goal of profit-making, which supersedes altruistic academic motivations as the ultimate goal of industry research labs [154].

A focus on accuracy in computer vision above all other measures of success, such as data or model efficiency, encourages indiscriminate collection of data. As Dotan and Milli write on values influencing the machine learning research community, "This kind of evaluation furthers certain values, such as centralization of power, while hindering other values, such as environmental sustainability and privacy" [31]. It is clear that significant attention needs to be applied to the ethical considerations of computer vision, especially considering the applied nature of the field, which is often closer to commercial applications than other domains of artificial intelligence. However, when only eight of 1,467 papers accepted for publication at CVPR 2020 are in the category of "Fairness, Accountability, Transparency and Ethics in Vision",[4] a subject area introduced for the *first time* in 2020, it is clear that there is much work to do to prioritize this line of scholarship.

In dealing with the issue of bias in datasets, one practical approach is to simply obtain more data of underrepresented groups. However, as marginalized communities are often the ones left out of datasets, increased collection intersects with issues of privacy and consent. Reports in 2019 of predatory actions by Google contractors in targeting Black people to obtain facial scans, particularly homeless individuals as "they're the least likely to say anything to the media," is illustrative of an approach that does not consider other important ethical issues when trying to diversify datasets [104]. Another approach to tackle bias in datasets is with technical solutions. Research in the technical fairness community seeks to craft algorithms that ensure fair outcomes from biased data by adding constraints during training that promote specified definitions of fairness, such as group fairness or equalized odds [6]. While a necessary tool in developing responsible computer vision systems, work in this domain can quickly disconnect practitioners from the entrenched biases in the world that create biased datasets in the first place, leading them to emphasize bias as a purely technical problem to solve [65].

Issues of bias are pervasive and require more than just technical fixes. Such a sentiment was echoed by participants in a 2019 survey of industry machine learning practitioners [58]. Through their interviews, Holstein et al. identified the need to move beyond algorithms to ensure responsible machine learning development,

> Through our investigation, we identify a range of real-world needs that have

[3] https://ai.googleblog.com/2017/07/revisiting-unreasonable-effectiveness.html
[4] https://openaccess.thecvf.com/CVPR2020

> been neglected in the literature so far, as well as several areas of alignment. For example, while the fair ML literature has largely focused on 'de-biasing' methods and viewed the training data as fixed, most of our interviewees report that their teams consider data collection, rather than model development, as the most important place to intervene. [58]

Transparency into datasets provides a good start for improving norms in the community. Implementing standardized documentation in data collection is a simple but effective way to convey relevant details of the dataset to end-users and to slow down the collection process, providing time for practitioners to reflect on the core assumptions and decisions that can embed biases into data. *Datasheets for Datasets* [44], *Data Statements for Natural Language Processing* [8], and *Dataset Nutrition Labels*[5] are works from various machine learning domains that propose frameworks to achieve these goals.

Lastly, diversity is key to responsible data collection and use in computer vision. As it currently stands, the artificial intelligence research community has a significant diversity problem. Women have been estimated to make up only 12% of leading researchers in the field [120] and Black students in the United States and Canada comprised fewer than 1% of all computer science PhD graduates in 2019 [155]. Diversity and inclusion are not just ideals we should strive for as a society, but underrepresented individuals bring with them new view-points and lived experiences, which can critique normative practices, positively impacting the field. The introduction of the feminist notion of "intersectionality" by Joy Buolamwini in *Gender Shades* has been transformative in machine learning model auditing and is a great example of how diverse backgrounds can improve computer vision [13]. Intersectionality, a term coined by Kimberlé Crenshaw in 1990 to describe the interconnected nature of social categories such as gender and race that create overlapping and independent systems of discrimination, which cannot be fully captured by looking at the categories separately [24], is now fundamental to identifying disparate performance impacts of models by disaggregating results on benchmark datasets. Better-supported diversity and inclusion initiatives in both academic and industry settings, along with more diversity in senior engineering and leadership roles, are desperately needed.

Data collection is an complex enterprise. As Jo and Gebru write in their work studying data collection through the lens of archivists, there is a need for an entire interdisciplinary subfield of machine learning, focused on data gathering, sharing, annotation, ethics monitoring and record keeping, such that machine learning researchers are more cognizant and systematic in data collection, especially of sociocultural data [64]. Establishing such a field would take considerable effort, yet, it is my hope that this thesis has demonstrated that

[5] https://datanutrition.org/

the normative data collection and use practices in computer vision are untenable, and that the establishment of such a field is long overdue.

In summary, through the presentation of (i) a novel audit of age and gender bias in the ImageNet dataset, (ii) novel evidence to suggest state-of-the-art face recognition systems exhibit a differential identification accuracy for individuals, dependent on their inclusion in training datasets that are collected without their consent, and (iii) novel analysis of a foundational dataset in facial expression recognition through the lens of new work on the nature of emotional expression, suggesting an overhaul is needed in the domain's conventional label taxonomy, this thesis hopes to challenge researchers to reconsider normative data collection and use practices such that computer vision systems can be developed in a more thoughtful and responsible manner.

References

[1] ACLU. ACLU calls for moratorium on law and immigration enforcement use of facial recognition, 2018. https://www.aclu.org/press-releases/aclu-calls-moratorium-law-and-immigration-enforcement-use-Facial-Recognition.

[2] Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baro, Isabelle Guyon, and Rasmus Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *12th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 87–94. IEEE, 2017.

[3] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787, 2018.

[4] Ankan Bansal, Anirudh Nanduri, Carlos D Castillo, Rajeev Ranjan, and Rama Chellappa. UMDFACES: An annotated face dataset for training deep networks. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 464–473. IEEE, 2017.

[5] Gregory Barber. The viral app that labels you isn't quite what you think. *Wired*, Sept 2019.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. 2019. http://www.fairmlbook.org.

[7] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[8] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018.

[9] Ruha Benjamin. 2020 vision: Reimagining the default settings of technology & society. In *The International Conference on Learning Representations, Invited Speaker*, 2020.

[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

[11] Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Larry D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 77–82. IEEE, 1994.

[12] Owen Bowcott. Police face legal action over use of facial recognition cameras. *The Guardian*, Jun 2018.

[13] 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.

[14] Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. http://www.pre.ethics.gc.ca/eng/documents/tcps2-2018-en-interactive-final.pdf, December 2018.

[15] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGDFACE2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[16] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 539–546, 2005.

[17] Albert Clapés, Ozan Bilici, Dariia Temirova, Egils Avots, Gholamreza Anbarjafari, and Sergio Escalera. From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2373–2382, 2018.

[18] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[19] Timothy Colburn and Gary Shute. Abstraction in computer science. *Minds and Machines*, 17(2):169–184, 2007.

[20] Kate Conger, Richard Fausset, and Serge F. Kovaleski. San Francisco Bans Facial Recognition Technology. *The New York Times*, May 2019.

[21] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.

[22] Kate Crawford, Roel Dobbe, Theodora Dryer, Genevieve Fried, Ben Green, Elizabeth Kaziunas, Amba Kak, Varoon Mathur, Erin McElroy, Andrea Nill Sánchez, et al. Ai now 2019 report. *New York, NY: AI Now Institute*, 2019.

[23] Kate Crawford and Trevor Paglen. Excavating AI: The Politics of Training Sets for Machine Learning. https://excavating.ai, September 2019.

[24] Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241, 1990.

[25] Gerard De Melo and Gerhard Weikum. Towards a universal wordnet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 513–522. ACM, 2009.

[26] Jia Deng. Large scale visual recognition. Technical report, Princeton University Department of Computer Science, 2012.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[28] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.

[29] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[30] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[31] Ravit Dotan and Smitha Milli. Value-laden disciplinary shifts in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 294–294, 2020.

[32] Chris Dulhanty and Ayush Sahni. Face recognition regulation in a global context. *CS 798-001*, Dec 2019.

[33] 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.

[34] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[35] Paul Ekman and Wallace V Friesen. Unmasking the face: A guide to recognizing emotions from facial clues. 1975.

[36] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[37] Li Fei-Fei. ImageNet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, volume 16, pages 18–25, 2010.

[38] Li Fei-Fei and Jia Deng. ImageNet: Where are we going? and where have we been. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[39] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–178, 2004.

[40] Thomas B Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

[41] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

[42] Timnit Gebru. Oxford handbook on ai ethics book chapter on race and gender. *arXiv preprint arXiv:1908.06165*, 2019.

[43] Timnit Gebru and Emily Denton. Tutorial on Fairness, Accountability, Transparency and Ethics in Computer Vision at CVPR 2020. https://sites.google.com/view/fatecv-tutorial/home, June 2020.

[44] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[45] Xin Geng and Rongzi Ji. Label distribution learning. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops*, pages 377–383, 2013.

[46] Darshan Gera and S Balasubramanian. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. *arXiv preprint arXiv:2007.10298*, 2020.

[47] Dave Gershgorn. The data that transformed ai research—and possibly the world. *Quartz*, 26:2017, 2017.

[48] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 87–102. Springer, 2016.

[49] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2018.

[50] Hu Han and Anil K. Jain. Age, gender and race estimation from unconstrained face images. Technical Report MSU-CSE-14-5, Michigan State University, 2014.

[51] Donna Haraway. *Simians, cyborgs, and women: The reinvention of nature*. Routledge, 2013.

[52] Adam Harvey and Jules LaPlace. Megapixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets, 2019.

[53] Drew Harwell. A face-scanning algorithm increasingly decides whether you deserve the job. *The Washington Post*, Nov 2019.

[54] 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.

[55] Kashmir Hill. Wrongfully accused by an algorithm. *The New York Times*, June 2020.

[56] Kashmir Hill and Aaron Krolik. How photos of your kids are powering surveillance technology. *The New York Times*, Oct 2019.

[57] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[58] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.

[59] Gary B. Huang, Honglak Lee, and Erik Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2518–2525, 2012.

[60] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[61] Illinois General Assembly. 740 ILCS 14 / Biometric Information Privacy Act, Oct 2008. http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

[62] Vidit Jain and Erik Learned-Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts, Amherst, 2010.

[63] Chi Jin, Ruochun Jin, Kai Chen, and Yong Dou. A community detection approach to cleaning extremely large face database. *Computational Intelligence and Neuroscience*, 2018, 2018.

[64] Eun Seo Jo and Timnit Gebru. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316, 2020.

[65] Khari Johnson. Ai weekly: A deep learning pioneer's teachable moment on ai bias. *VentureBeat*, June 2020.

[66] 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*, pages 3819–3828. ACM, 2015.

[67] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[68] Nicolas Kayser-Bril. Google apologizes after its vision ai produced racist results. *AlgorithmWatch*, April 2020.

[69] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[70] 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.

[71] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[73] Erik Learned-Miller, Gary B. Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.

[74] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10143–10152, 2019.

[75] Fei-Fei Li, Justin Johnson, and Serena Yeung. CS231n: Convolutional Neural Networks for Visual Recognition, Spring 2018, Lecutre 10. http://cs231n.stanford.edu/slides/2018/cs231n_2018_lecture10.pdf, May 2018.

[76] Pei Li and Adam Jourdan. Sleepy pupils in the picture at high-tech chinese school. *Reuters*, May 2018.

[77] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 2020.

[78] J Robert Lilly, Francis T Cullen, and Richard A Ball. *Criminological theory: Context and consequences*. Sage publications, 2018.

[79] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[80] Anthony C Little, Benedict C Jones, and Lisa M DeBruine. Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1571):1638–1659, 2011.

[81] Ping Liu, Yuewei Lin, Zibo Meng, Weihong Deng, Joey Tianyi Zhou, and Yi Yang. Point adversarial self mining: A simple method for facial expression recognition in the wild. *arXiv preprint arXiv:2008.11401*, 2020.

[82] 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*, pages 212–220, 2017.

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

[84] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205. IEEE, 1998.

[85] Michael J Lyons, Shigeru Akamatsu, Miyuki Kamachi, Jiro Gyoba, and Julien Budynek. The japanese female facial expression (JAFFE) database. In *Proceedings of Third International Conference on Automatic Face and Gesture Recognition*, pages 14–16, 1998.

[86] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision*, pages 181–196, 2018.

[87] John Markoff. Seeking a better way to find web images. *The New York Times*, 2012.

[88] Dimitrios Marmanis, Mihai Datcu, Thomas Esch, and Uwe Stilla. Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105–109, 2015.

[89] Yunliang Meng. Racially biased policing and neighborhood characteristics: A case study in Toronto, Canada. *Cybergeo: European Journal of Geography*, 2014.

[90] Ryan Merkley. Use and fair use: Statement on shared images in facial recognition ai, Mar 2019.

[91] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[92] Jacob Metcalf and Kate Crawford. Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211, 2016.

[93] Cade Metz. 'nerd,' 'nonsmoker,' 'wrongdoer': How might a.i. label you? *The New York Times*, Sept 2019.

[94] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[95] Clayton James Mosher. *Discrimination and denial: Systemic racism in Ontario's legal and criminal justice systems, 1892-1961*. University of Toronto Press, 1998.

[96] Paul Mozur. In Hong Kong protests, faces become weapons. *The New York Times*, Jul 2019.

[97] Madhumita Murgia. Microsoft quietly deletes largest public face recognition data set. *Financial Times*, Jun 2019.

[98] NAACP. Criminal justice fact sheet, 2018. http://www.naacp.org/criminal-justice-fact-sheet/.

[99] Aaron Nech and Ira Kemelmacher-Shlizerman. Level playing field for million scale face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[100] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (COIL-20).

[101] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 343–347. IEEE, 2014.

[102] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

[103] George Orwell. *1984*. Secker & Warburg, London, UK, 1949.

[104] Ginger Adams Otis and Nancy Dillon. Google using dubious tactics to target people with 'darker skin' in facial recognition project: sources. *New York Daily News*, Oct 2019.

[105] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[106] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[107] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[108] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 429–435, 2019.

[109] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Medicine*, 15(11):e1002686, 2018.

[110] Sarah Ravani. Oakland bans use of facial recognition technology, citing bias concerns. *San Francisco Chronicle*, Jul 2019.

[111] Kevin Rector and Alison Knezevich. Maryland's use of facial recognition software questioned by researchers, civil liberties advocates. *The Baltimore Sun*, Oct 2016.

[112] Lauren Rhue. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*, 2018.

[113] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision workshop on Benchmarking Multi-Target Tracking*, 2016.

[114] Claudia Veronica Roberts. Quantifying the extent to which popular pre-trained convolutional neural networks implicitly learn high-level protected attributes. Master's thesis, Princeton University, 2018.

[115] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[116] 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.

[117] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[118] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015.

[119] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 Workshop: Machine Learning for the Developing World*, 2017.

[120] Tom Simonite. Ai is the future—but where are the women? *Wired*, Aug 2018.

[121] Norma Smith, Carolyn Eslick, Sharon Tomiko Santos, Gerry Pollet, and Shelley Kloba. HB 2644 - 2019-20, Concerning Artificial Intelligence-Enabled Profiling. https://app.leg.wa.gov/billsummary?BillNumber=2644&Year=2019.

[122] Olivia Solon. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. *NBCNews.com*, Mar 2019.

[123] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2325–2333, 2016.

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

[125] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 843–852, 2017.

[126] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[127] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[128] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[129] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2746–2754, 2015.

[130] Daniel Thomas. The cameras that know if you're happy - or a threat. *BBC*, Jul 2018.

[131] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[132] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[133] 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*, pages 5265–5274, 2018.

[134] Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.

[135] Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[136] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[137] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Myers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. AI Now Report 2018. *AI Now Institute*, 2018.

[138] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[139] Lior Wolf, Tal Hassner, and Itay Maoz. *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011.

[140] Sarah Wu. Somerville city council passes facial recognition ban. *The Boston Globe*, Jun 2019.

[141] Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images.

[142] Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. SCUT-FBP: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1821–1826. IEEE, 2015.

[143] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves ImageNet classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[144] Blaise Aguera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy's new clothes. *Medium (6 May 2017), online:¡ https://medium. com/@ blaisea/physiognomys-new-clothesf2d4b59fdd6a*, 2017.

[145] 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.

[146] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5525–5533, 2016.

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

[148] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[149] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[150] Ling Zhang, Le Lu, Isabella Nogues, Ronald M Summers, Shaoxiong Liu, and Jianhua Yao. Deeppap: deep convolutional networks for cervical cell classification. *IEEE Journal of Biomedical and Health Informatics*, 21(6):1633–1643, 2017.

[151] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *IEEE International Joint Conference on Biometrics*, pages 1–9. IEEE, 2017.

[152] Ting Zhang. Facial expression recognition based on deep learning: a survey. In *International Conference on Intelligent and Interactive Systems and Applications*, pages 345–352. Springer, 2017.

[153] Ying Zhou, Hui Xue, and Xin Geng. Emotion distribution recognition from facial expressions. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1247–1250, 2015.

[154] Shoshana Zuboff. The age of surveillance capitalism, 2019.

[155] Stuart Zweben and Betsy Bizot. 2019 taulbe survey: Total undergrad cs enrollment rises again, but with fewer new majors; doctoral degree production recovers from last year's dip. *Computing Research Association*, May 2020.

Appendices

Appendix A

Categorical JAFFE Citations — January 1 to June 30 2020

[1] Rim Afdhal, Ridha Ejbali, and Mourad Zaied. Emotion recognition by a hybrid system based on the features of distances and the shapes of the wrinkles. *The Computer Journal*, 63(3):351–363, 2020.

[2] Khaled M Alalayah, Reyazur Rashid Irashad, Taha H Rassem, and Badiea Abdulkarem Mohammed. A new fast local laplacian completed local ternary count (fll-cltc) for facial image classification. *IEEE Access*, 2020.

[3] Ying Bi, Bing Xue, and Mengjie Zhang. An effective feature learning approach using genetic programming with image descriptors for image classification [research frontier]. *IEEE Computational Intelligence Magazine*, 15(2):65–77, 2020.

[4] Ying Bi, Bing Xue, and Mengjie Zhang. Genetic programming with a new representation to automatically learn features and evolve ensembles for image classification. *IEEE Transactions on Cybernetics*, 2020.

[5] Yonghe Chu, Hongfei Lin, Liang Yang, Yufeng Diao, Dongyu Zhang, Shaowu Zhang, Xiaochao Fan, Chen Shen, Bo Xu, and Deqin Yan. Discriminative globality-locality preserving extreme learning machine for image classification. *Neurocomputing*, 387:13–21, 2020.

[6] Paramartha Dutta and Asit Barman. Distance-texture signature duo for determination of human emotion. In *Human Emotion Recognition from Face Images*, pages 125–176. Springer, 2020.

[7] Paramartha Dutta and Asit Barman. Human emotion recognition using combination of shape-texture signature feature. In *Human Emotion Recognition from Face Images*, pages 177–234. Springer, 2020.

[8] Kushal Kanti Ghosh, Ritam Guha, Soulib Ghosh, Suman Kumar Bera, and Ram Sarkar. Atom search optimization with simulated annealing–a hybrid metaheuristic approach for feature selection. *arXiv preprint arXiv:2005.08642*, 2020.

[9] Sonia M González-Lozoya, Jorge de la Calleja, Luis Pellegrin, Hugo Jair Escalante, Ma Auxilio Medina, and Antonio Benitez-Ruiz. Recognition of facial expressions based on cnn features. *Multimedia Tools and Applications*, pages 1–21, 2020.

[10] Mahesh Goyani and Narendra Patel. Template matching and machine learning-based robust facial expression recognition system using multi-level haar wavelet. *International Journal of Computers and Applications*, 42(4):360–371, 2020.

[11] Shasha Guo, Lianhua Qu, Lei Wang, Xulong Tang, Shuo Tian, Shiming Li, and Weixia Xu. Exploration of input patterns for enhancing the performance of liquid state machines. *arXiv preprint arXiv:2004.02540*, 2020.

[12] Ahmed Rachid Hazourli, Amine Djeghri, Hanan Salam, and Alice Othmani. Deep multi-facial patches aggregation network for facial expression recognition. *arXiv preprint arXiv:2002.09298*, 2020.

[13] Koffi Eddy Ihou and Nizar Bouguila. Stochastic topic models for large scale and nonstationary data. *Engineering Applications of Artificial Intelligence*, 88:103364, 2020.

[14] Shruti Jaiswal and Gora Chand Nandi. Hyperparameters optimization for deep learning based emotion prediction for human robot interaction. *arXiv preprint arXiv:2001.03855*, 2020.

[15] Kapil Juneja and Chhavi Rana. Multi-featured and fuzzy-filtered machine learning model for face expression classification. *Wireless Personal Communications*, pages 1–30, 2020.

[16] Jenni Kommineni, Satria Mandala, Mohd Shahrizal Sunar, and Parvathaneni Midhu Chakravarthy. Advances in computer–human interaction for detecting facial expression using dual tree multi band wavelet transform and gaussian mixture model. *Neural Computing and Applications*, pages 1–12, 2020.

[17] Hyun-Soon Lee and Bo-Yeong Kang. Continuous emotion estimation of facial expressions on jaffe and ck+ datasets for human–robot interaction. *Intelligent Service Robotics*, pages 1–13, 2020.

[18] Shan Li and Weihong Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 2020.

[19] Yangyang Li, Shuangkang Fang, Xiaoyu Bai, Licheng Jiao, and Naresh Marturi. Parallel design of sparse deep belief network with multi-objective optimization. *Information Sciences*, 2020.

[20] Ping Liu, Yunchao Wei, Zibo Meng, Weihong Deng, Joey Tianyi Zhou, and Yi Yang. Omni-supervised facial expression recognition: A simple baseline. *arXiv preprint arXiv:2005.08551*, 2020.

[21] V Uma Maheswari, G Varaprasad, and S Viswanadha Raju. Local directional maximum edge patterns for facial expression recognition. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–9, 2020.

[22] Hemant Kumar Meena, Kamalesh Kumar Sharma, and Shiv Dutt Joshi. Effective curvelet-based facial expression recognition using graph signal processing. *Signal, Image and Video Processing*, 14(2):241–247, 2020.

[23] Avishek Nandi, Paramartha Dutta, and Md Nasir. Recognizing human emotions from facial images by landmark triangulation: A combined circumcenter-incenter-centroid trio feature-based method. In *Algorithms in Machine Learning Paradigms*, pages 147–164. Springer, 2020.

[24] Vansh Narula, Theodora Chaspari, et al. An adversarial learning framework for preserving users' anonymity in face-based emotion recognition. *arXiv preprint arXiv:2001.06103*, 2020.

[25] Seong-Gi Oh and TaeYong Kim. Facial expression recognition by regional weighting with approximated q-learning. *Symmetry*, 12(2):319, 2020.

[26] Shreya Pendsey, Eshani Pendsey, and Shweta Paranjape. Empathic diary based on emotion recognition using convolutional neural network. In *Proceeding of International Conference on Computational Science and Applications*, pages 3–14. Springer, 2020.

[27] Chong Peng, Zhilu Zhang, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. Two-dimensional semi-nonnegative matrix factorization for clustering. *arXiv preprint arXiv:2005.09229*, 2020.

[28] Nazil Perveen and Chalavadi Krishna Mohan. Configural representation of facial action units for spontaneous facial expression recognition in the wild. In *VISIGRAPP (4: VISAPP)*, pages 93–102, 2020.

[29] Shuvendu Roy. Island loss for improving the classification of facial attributes with transfer learning on deep convolutional neural network. *International Journal of Image, Graphics and Signal Processing*, 12(1):18, 2020.

[30] Soumyajit Saha, Manosij Ghosh, Soulib Ghosh, Shibaprasad Sen, Pawan Kumar Singh, Zong Woo Geem, and Ram Sarkar. Feature selection for facial emotion recognition using cosine similarity-based harmony search algorithm. *Applied Sciences*, 10(8):2816, 2020.

[31] Sumeet Saurav, Sanjay Singh, Ravi Saini, and Madhulika Yadav. Facial expression recognition using improved adaptive local ternary pattern. In *Proceedings of 3rd International Conference on Computer Vision and Image Processing*, pages 39–52. Springer, 2020.

[32] Natalya Selitskaya, S Sielicki, Livija Jakaite, Vitaly Schetinin, F Evans, Marc Conrad, and Paul Sant. Deep learning for biometric face recognition: Experimental study on benchmark data sets. In *Deep Biometrics*, pages 71–97. Springer, 2020.

[33] Jacopo Sini, Antonio Costantino Marceddu, and Massimo Violante. Automatic emotion recognition for the calibration of autonomous driving functions. *Electronics*, 9(3):518, 2020.

[34] A Swaminathan, A Vadivel, and Michael Arock. Ferce: Facial expression recognition for combined emotions using ferce algorithm. *IETE Journal of Research*, pages 1–16, 2020.

[35] Kamlesh Tiwari and Mayank Patel. Facial expression recognition using random forest classifier. In *International Conference on Artificial Intelligence: Advances and Applications 2019*, pages 121–130. Springer, 2020.

[36] Sabyasachi Tribedi and Ranjit Kumar Barai. Generating context-free group-level emotion landscapes using image processing and shallow convolutional neural networks. In *Progress in Computing, Analytics and Networking*, pages 313–325. Springer, 2020.

[37] Tuo Wang, Xiang Zhang, Long Lan, Qing Liao, Chuanfu Xu, and Zhigang Luo. Correlation self-expression shrunk for subspace clustering. *IEEE Access*, 8:16595–16605, 2020.

[38] Wenxuan Wang, Yanwei Fu, Qiang Sun, Tao Chen, Chenjie Cao, Ziqi Zheng, Guoqiang Xu, Han Qiu, Yu-Gang Jiang, and Xiangyang Xue. Learning to augment expressions for few-shot fine-grained facial expression recognition. *arXiv preprint arXiv:2001.06144*, 2020.

[39] Yingying Wang, Yibin Li, Yong Song, and Xuewen Rong. The influence of the activation function in a convolution neural network model of facial expression recognition. *Applied Sciences*, 10(5):1897, 2020.

[40] Zheng Wang, Lin Zuo, Jing Ma, Si Chen, Jingjing Li, Zhao Kang, and Lei Zhang. Exploring nonnegative and low-rank correlation for noise-resistant spectral clustering. *World Wide Web*, pages 1–21, 2020.

[41] Xiang-Fei Yang, Yuan-Hai Shao, Chun-Na Li, Li-Ming Liu, and Nai-Yang Deng. Principal component analysis based on t-norm maximization. *arXiv preprint arXiv:2005.12263*, 2020.

[42] Guanghao Zhang, Dongshun Cui, Shangbo Mao, and Guang-Bin Huang. Unsupervised feature learning with sparse bayesian auto-encoding based extreme learning machine. *International Journal of Machine Learning and Cybernetics*, pages 1–13, 2020.

[43] Guanghao Zhang, Yue Li, Dongshun Cui, Shangbo Mao, and Guang-Bin Huang. R-elmnet: Regularized extreme learning machine network. *Neural Networks*, 2020.

[44] Hepeng Zhang, Bin Huang, and Guohui Tian. Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. *Pattern Recognition Letters*, 131:128–134, 2020.

[45] Lingling Zhang, Minnan Luo, Jun Liu, Zhihui Li, and Qinghua Zheng. Diverse fuzzy c-means for image clustering. *Pattern Recognition Letters*, 130:275–283, 2020.

[46] Wei Zheng, Xiaofeng Zhu, Guoqiu Wen, Yonghua Zhu, Hao Yu, and Jiangzhang Gan. Unsupervised feature selection by self-paced learning regularization. *Pattern Recognition Letters*, 132:4–11, 2020.

Appendix B

Scored JAFFE Citations — January 1 to June 30 2020

[1] Abeer Almowallad and Victor Sanchez. Human emotion distribution learning from face images using cnn and lbc features. In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2020.

[2] Manuel González, José-Ramón Cano, and Salvador García. Prolsfeo-ldl: Prototype selection and label-specific feature evolutionary optimization for label distribution learning. *Applied Sciences*, 10(9):3089, 2020.

[3] Sebastian Mathias Keller, Maxim Samarin, Fabricio Arend Torres, Mario Wieser, and Volker Roth. Learning extremal representations with deep archetypal analysis. *arXiv preprint arXiv:2002.00815*, 2020.

[4] Suping Xu, Hengrong Ju, Lin Shang, Witold Pedrycz, Xibei Yang, and Chun Li. Label distribution learning: A local collaborative mechanism. *International Journal of Approximate Reasoning*, 121:59–84, 2020.