Mitigating dataset harms requires stewardship: Lessons from 1000 papers

Kenny Peng, Arunesh Mathur, Arvind Narayanan
Princeton University

Draft: August 9, 2021

Abstract

Concerns about privacy, bias, and harmful applications have shone a light on the ethics of machine learning datasets, even leading to the retraction of prominent datasets including DukeMTMC, MS-Celeb-1M, TinyImages, and VGGFace2. In response, the machine learning community has called for higher ethical standards, transparency efforts, and technical fixes in the dataset creation process. The premise of our work is that these efforts can be more effective if informed by an understanding of how datasets are used in practice in the research community. We study three influential face and person recognition datasets—DukeMTMC, MS-Celeb-1M, and Labeled Faces in the Wild (LFW)—by analyzing nearly 1000 papers that cite them. We found that the creation of derivative datasets and models, broader technological and social change, the lack of clarity of licenses, and dataset management practices can introduce a wide range of ethical concerns. We conclude by suggesting a distributed approach that can mitigate these harms, making recommendations to dataset creators, conference program committees, dataset users, and the broader research community.

1 Introduction

Datasets play an essential role in machine learning research, but are also the source of ethical concerns. These include concerns about the privacy of individuals included in datasets [36, 62], representational harms introduced by annotations [20, 35], effects of biases in datasets on downstream use [16, 17, 15], and the use of datasets for ethically dubious purposes [36, 67, 59]. These concerns have led to the retractions of several prominent research datasets including Tiny Images, VGGFace2, DukeMTMC, and MS-Celeb-1M.

Mitigating harms associated with datasets is now recognized as an important goal by the machine learning community. Researchers have worked to make sense of ethical considerations involved in dataset creation [34, 61, 25], and have proposed ways to identify and mitigate biases in datasets [8, 73], protect the privacy of individuals [62, 81], and document datasets [28, 39, 9, 58].

The premise of our work is that these efforts can be more effective if informed by an understanding of how datasets are used in practice in the research community. We present an account of the life cycles of three popular face and person recognition datasets: Labeled Faces in the Wild (LFW) [42], MS-Celeb-1M [32], and DukeMTMC [65]. We selected these because "people-centric" datasets [34] have been the subject of especially serious ethical concerns. We analyze nearly 1000 papers that cite these datasets and their derivative datasets and pre-trained models (Section 2). We present five primary findings:

- We reveal limitations of dataset retraction in mitigating harms (Section 3). We find that after the retractions of DukeMTMC and MS-Celeb-1M, the underlying data of both datasets remained widely available. Both datasets were used hundreds of times in papers published months after the retractions, possibly due to a lack of clarity in both retractions. Because of such "runaway data," retractions are unlikely to cut off data access; moreover, without a clear indication of intentions, retractions may have limited normative influence.
- We show how derivatives raise new ethical considerations (Section 4). Across DukeMTMC, MS-Celeb-1M, and LFW, we identified 35 derivative datasets and six classes of pre-trained models. We document four ways in which derivatives can cause ethical concerns: by enabling new applications, enabling use of the dataset in production settings, introducing new annotations of the data, or applying additional post-processing such as cleaning. Thus, the impact of a dataset may be much broader than its original intention.

- We show how the ethical concerns associated with a dataset can change over time, both as a result of technological and social change (Section 5). In the case of LFW and the influential ImageNet dataset, technological advances opened the door for production use of the datasets, raising new ethical concerns. Additionally, various social factors led to a more critical understanding of the demographic composition of LFW and the annotation practices underlying ImageNet.

- We show how licenses, a primary mechanism governing dataset use, can lack substantive effect (Section 6). We find that the licenses of DukeMTMC, MS-Celeb-1M, and LFW do not effectively restrict production use of the datasets. In particular, while the original license of MS-Celeb-1M indicates only non-commercial research use of the dataset, 19 of 21 GitHub repositories we found containing models pre-trained on MS-Celeb-1M included such a designation. We find anecdotal evidence suggesting that production use of models trained on non-commercial datasets is commonplace.

- We show that while dataset management and citation practices can support harm mitigation, current practices have several shortcomings (Section 7). Specifically, we find that dataset documentation is not easily accessible from citations and is not persistent. Moreover, dataset use is not clearly specified in academic papers, often resulting in ambiguities. Finally, current infrastructure does not support the tracking of dataset use or of derivatives in order to retrospectively understand the impact of datasets.

In short, we found that the creation of derivative datasets and models, broader technological and social change, the lack of clarity of licenses, and dataset management practices can introduce a wide range of ethical concerns. Based on these findings, we conclude the paper with recommendations for harm mitigation in machine learning research. Our approach emphasizes steps that can be taken after dataset creation, which we call dataset stewarding. We advocate for responsibility to be distributed among many stakeholders including dataset creators, conference program committees, dataset users, and the broader research community.

2 Overview of the datasets and analysis

We analyzed the life cycles of three popular face and person recognition datasets: DukeMTMC [65], MS-Celeb-1M [32], and Labeled Faces in the Wild (LFW) [42]. We refer to these datasets as "parent datasets." To develop a fuller understanding of the impact of each parent dataset, we also aimed to capture the use of their derivatives (Figure 1). We provide a summary of our methodology below.

We constructed a corpus of papers that potentially used each parent dataset or their derivatives. To do this, we first compiled a list of derivatives of each parent dataset. Then we found the papers closely associated with each parent and derived dataset. Finally, we compiled lists of papers citing each associated paper using Semantic Scholar [27] (see Table 1). One coder then reviewed a sample of these papers, indicating if a paper used the parent dataset or a derivative. In total, our analysis includes 951 papers (275 citing DukeMTMC or its derivatives, 276 citing MS-Celeb-1M or its derivatives, and 400 citing LFW or its derivatives). We found many papers that used derivatives that were not included in our list, and suspect that other derivatives exist. In general, our results should be viewed as "lower bounds."

The three datasets we analyze here are each popular face recognition or person recognition datasets. After collecting a list of 54 datasets, we chose LFW, which was the most cited dataset in our list. Since LFW was introduced in 2007, it allowed us to perform a longitudinal analysis. We then chose DukeMTMC and MS-Celeb-1M, which were the two most cited datasets in our list that had been retracted. (VGGFace2, another dataset on our list, was retracted after we began our analysis.) These two datasets allowed us to study the effects of retraction.

We now provide background information for the three datasets:

- **MS-Celeb-1M** was introduced by Microsoft researchers in 2016 as a face recognition dataset [32]. It includes about 10 million images of about 10,000 "celebrities." The original paper gave no specific motivating applications, but did note that "Recognizing celebrities, rather than a pre-selected private group of people, represents public interest and could be directly applied to a wide range of real scenarios." Researchers and journalists noted in 2019 that many of the "celebrities" were in fact fairly ordinary citizens, and that the images were aggregated without consent [36, 59]. Several corporations tied to mass surveillance operations were also found to use the dataset in research papers [36, 59]. The dataset was taken down in June 2019. Microsoft, in a statement to the Financial Times, said that the reason was "because the research challenge is over." [59]

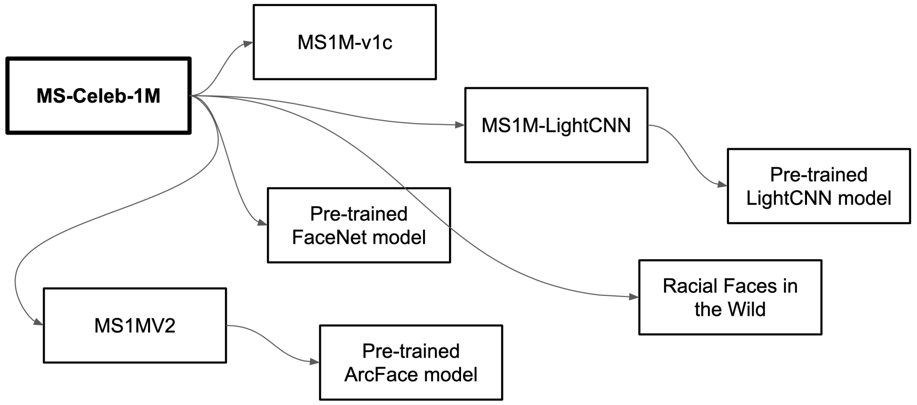
MS-Celeb-1M
MS1M-v1c
MS1M-LightCNN
Pre-trained LightCNN model
Pre-trained FaceNet model
Racial Faces in the Wild
MS1MV2
Pre-trained ArcFace model

Figure 1: The impact of a dataset extends beyond its direct use. Thus, our analysis includes the use of its derivatives. Here, some of MS-Celeb-1M's derivatives are shown.

- **DukeMTMC** was introduced in 2016 as a benchmark for evaluating multi-target multi-camera tracking systems, which "automatically track multiple people through a network of cameras." [65] The dataset's creators defined performance measures aimed at applications where preserving identity is important, such as "sports, security, or surveillance." The images were collected from video footage taken on Duke's campus. The same reports on MS-Celeb-1M listed above [36, 59] noted that the DukeMTMC was also being used by corporations tied to mass surveillance operations, and also noted the lack of consent given by people included in the dataset. The creators removed the dataset in June 2019, apologizing, noting that they had inadvertently broken guidelines provided by the Duke University IRB.
- **LFW** was introduced in 2007 as a benchmark dataset for face verification [42]. It was one of the first face recognition datasets that included faces from an unconstrained "in-the-wild" setting, using faces scraped from Yahoo News articles (via the Faces in the Wild dataset [12]). In the originally-released paper, the dataset's creators gave no motivating applications or intended uses beyond studying face recognition. In fall 2019, a disclaimer was added to the dataset's associated website, noting that the dataset should not be used to "conclude that an algorithm is suitable for any commercial purpose." [3]

Dataset id	Dataset name	Associated paper	Dataset or model	Assoc. paper sampled	Num. sampled	Doc. uses	2020 doc. uses	New application	Attribute annotations	Post-processing	Still available	Includes orig. imgs.	License type	Prohibits comm. use
D1	DukeMTMC	[65]	dataset	✓	164	14	1					✓	CC BY-NC-SA 4.0	✓
D2	DukeMTMC-ReID	[84]	dataset	✓	172	142	63	✓			✓	✓	CC BY-NC-SA 4.0	✓
D3	DukeMTMC-VideoReID	[78]	dataset	✓	24	11	5	✓			✓	✓	CC BY-NC-SA 4.0	✓
D4	DukeMTMC-Attribute	[49]	dataset			10	1	✓	✓		✓		CC BY-NC-SA 4.0	✓
D5	DukeMTMC4ReID	[29]	dataset			3	0	✓				✓	CC BY-NC-SA 4.0	✓
D6*	DukeMTMC Group	[79]	dataset			3	1	✓						
D7	DukeMTMC-SI-Tracklet	[46]	dataset			1	1	✓				✓	CC BY-NC-SA 4.0	✓
D8	Occluded-DukeMTMC	[56]	dataset			1	1	✓			✓[1]	✓	CC BY-NC-SA 4.0	✓
D9	MS-Celeb-1M	[32]	dataset	✓	153	41	11				✓	✓	custom,[2] none	✓
D10	MS1MV2	[22]	dataset	✓	183	13	8			✓	✓	✓	none	
D11	MS1M-RetinaFace	[23]	dataset			2	2			✓	✓	✓	none	
D12	MS1M-LightCNN	[77]	dataset			3	0			✓	✓		MIT	
D13	MS1M-IBUG	[24]	dataset			3	1			✓	✓	✓	none	
D14	MS-Celeb-1M-v1c	[4]	dataset			6	4			✓	✓	✓	custom	✓
D15	RFW	[74]	dataset			1	1	✓	✓		✓	✓	custom	✓
D16	MS-Celeb-1M lowshot	[32]	dataset			4	0					✓	custom	✓
D17*	Universe	[6]	dataset			2	1							
M1	VGGFace		model			6	3	✓			✓		MIT, none	
M2	Prob. Face Embeddings		model			1	1	✓			✓		MIT	
M3	ArcFace / InsightFace		model			14	13	✓			✓		MIT, Apache 2.0, custom, none	some
M4	LightCNN		model			4	3	✓			✓		custom, none	some
M5	FaceNet		model			12	5	✓			✓		MIT, none	
M6	DREAM		model			1	1	✓			✓		BSD-2-Clause	
D18	LFW	[42]	dataset	✓	220	105					✓	✓	none	
D19	LFWA	[50]	dataset	✓	158	2		✓	✓		✓	✓	none	
D20	LFW-a	[76]	dataset	✓	31	14				✓	✓	✓	none	
D21	LFW3D	[37]	dataset	✓	24	3				✓	✓	✓	none	
D22	LFW deep funneled	[41]	dataset	✓	18	4				✓	✓	✓	none	
D23	LFW crop	[66]	dataset	✓	8	2				✓	✓	✓	none	
D24	BLUFR protocol	[48]	dataset		2	1					✓		none	
D25*	LFW87	[47]	dataset	✓	7	1								
D26	LFW+	[33]	dataset	✓	12	0		✓	✓		✓	✓	custom	✓
D27	<no name given>	[45]	dataset			4		✓	✓		✓		none	
D28	<no name given>	[31]	dataset			4					✓		none	
D29	SMFRD	[82]	dataset			1		✓			✓	✓	none	
D30	LFW funneled	[40]	dataset			2				✓	✓	✓	none	
D31	<no name given>	[2]	dataset			2		✓	✓				none	
D32	<no name given>	[13]	dataset			1					✓		none	
D33	MTFL	[83]	dataset			1		✓	✓		✓	✓	none	
D34	PubFig83 + LFW	[7]	dataset			2					✓	✓	none	
D35	Front. Faces in the Wild	[26]	dataset			1					✓	✓	none	
D36	ITWE	[85]	dataset			1		✓			✓	✓	custom	✓
D37	Extended LFW	[70]	dataset			2					✓	✓	none	
D38	<no name given>	[21]	dataset			1							custom	✓

Table 1: Summary of our overarching analysis.

Condensed key for Table 1. *assoc. paper sampled* — yes if our corpus included a sample of papers citing the dataset's associated paper(s); *doc. uses* — the number of uses of the dataset that we were able to document; *new application* — if the derivative explicitly or implicitly enables a new application that can raise ethical questions; *attribute annotation* — if the derivative includes labels for sensitive attributes such as race or gender; *post-processing* — if the derivative manipulates the original images (for example, by cleaning or aligning); *prohibits comm. use* — if the dataset or model's license information includes a non-commercial clause; in *dataset id*, an asterisk (*) indicates that we were unable to identify where the dataset is or was made available; in *dataset name*, some datasets were not given names by their creators; in *license type*, we give multiple licenses when the dataset or derivative is made available in multiple locations.

[1] The dataset itself is no longer available. However, a script to convert DukeMTMC-ReID (which is still available) to Occluded-DukeMTMC remains available.

[2] The original MS-Celeb-1M license is no longer publicly accessible. An archived version is available at `http://web.archive.org/web/20170430224804/http://msceleb.blob.core.windows.net/ms-celeb-v1-split/MSR_LA_Data_MSCeleb_IRC.pdf`.

3 Retractions and runaway data

When datasets are deemed problematic by the machine learning community, activists, or the media, dataset creators have responded by retracting them. MS-Celeb-1M [32], DukeMTMC [65], VGGFace [60], and Brainwash [69] were all retracted after an investigation by Harvey and Laplace [36] that pointed out ethical concerns with how the data was collected and being used. TinyImages [72] was retracted following a report by Prabhu and Birhane [62] that raised ethical concerns involving offensive labels in the dataset.

Retractions such as these may mitigate harm in two primary ways. First, they may place hard limitations on dataset use by making the data unavailable. Second, they may exert a normative influence, indicating to the community that the data should no longer be used. This can allow publication venues and other bodies to place their own limitations on such use.

With this in mind, we report a detailed analysis of the retractions of DukeMTMC and MS-Celeb-1M, and the effects of those retractions. We show that both fall short of effectively accomplishing either of the above mentioned goals. Both MS-Celeb-1M and DukeMTMC continue to be available for download through file sharing websites and through derivatives. Both datasets also continue to be used by the machine learning community in peer-reviewed research. The phenomenon where data is available through a multitude of sources outside a creator's control may be called "runaway data"—a term coined by Harvey and Laplace [36]. Finally, both retractions lacked specificity and clarity, which may have contributed to confusion in the community about whether it is ethically acceptable to continue to use them. We summarize these findings in Table 2 and elaborate below.

Continued availability. Despite their retractions in June 2019, data from MS-Celeb-1M and DukeMTMC remain publicly available. Five of the seven derivatives of DukeMTMC either contained subsets of or the entire original dataset. The two most popular derivatives—DukeMTMC-ReID [84] and DukeMTMC-VideoReID [78]—are still available for download to this day. Both DukeMTMC-ReID and DukeMTMC-VideoReID contain a cropped and edited subset of the videos from DukeMTMC.

Similarly, six derivatives of MS-Celeb-1M contained subsets of or the entire original dataset. Four of these—MS1MV2 [22], MS1M-RetinaFace [23], MS1M-IBUG [24], and MS-Celeb-1M-v1c [4]—are still available for download to this day. Racial Faces in the Wild [74] also appears available, but requires sending an email to obtain access. Further, we found that the original MS-Celeb-1M dataset, while taken down by Microsoft, continues to be available through third-party sites such as Academic Torrents [19]. We also identified 20 GitHub repositories that continue to make available models pre-trained on MS-Celeb-1M data.

Clearly, one of the goals of retraction is to limit the availability of datasets. Achieving this goal requires addressing all locations where the data might already be or might become available.

Continued use. Besides being available, both MS-Celeb-1M and DukeMTMC have been used in numerous research papers after they were retracted in June 2019. Within our sample of papers, we found that DukeMTMC and its derivatives had been used 73 times and MS-Celeb-1M and its derivatives had been used 54 times in 2020. Because our samples are 20% of our entire corpus, this equates to hundreds of uses in total. (See Figure 2 for a comparison of use to previous years.)

This use further highlights the limits of retraction. Many of the cases we identified involved derivatives that were not retracted. Indeed, 72 of 73 DukeMTMC uses were through derivative datasets, 63 of which came from the DukeMTMC-ReID dataset, a derivative that continued to be available. Similarly, only 11 of 54 MS-Celeb-1M uses were through the original dataset, while 17 were through derivative datasets and 26 were through pre-trained models.

One limitation of our analysis is that the use of a dataset in a paper published in 2020 (six months or more after retraction) could mean several things. The research could have been initiated after retraction, with the researchers ignoring the retraction and obtaining the data through a copy or a derivative. The research could have begun before the retraction and the researchers may not have learned of the retraction. Or, the research could already have been under review. Regardless, it is clear that 18 months after the retractions, they have not had the effect that one might have hoped for.

Retractions lacked specificity and clarity. In light of the continued availability and use of both these datasets, it is worth considering whether the retractions included sufficient information about why other researchers should refrain from using the dataset.

	DukeMTMC	**MS-Celeb-1M**
Availability of original	We did not find any locations where the original dataset is still available.	The dataset is still available through Academic Torrents and Archive.org.
Availability of derived datasets	We found two derived datasets that remain available and include the original images.	We found five derived datasets that remain available and include the original images.
Availability of pre-trained models	We did not find any models pre-trained on DukeMTMC data that are still available.	We found 20 GitHub repositories that still contain models pre-trained on MS-Celeb-1M data.
Continued use	In our sample, DukeMTMC and its derivatives were used 73 times in papers published in 2020.	In our sample, MS-Celeb-1M and its derivatives were used 54 times in papers published in 2020.
Status of original dataset page	The original website (`http://vision.cs.duke.edu/DukeMTMC/`) returns a DNS error.	The original website (`https://www.msceleb.org`) only contains filler text.
Other statements made by creators	A creator of DukeMTMC issued an apology, noting that the data collection violated IRB guidelines in two respects: "Recording outdoors rather than indoors, and making the data available without protections" [71].	In June 2019, Microsoft said that the dataset was taken down "because the research challenge is over" [59].
Availability of metadata	The license is no longer officially available, but is still available through GitHub repositories of derivative datasets.	The license is no longer officially available.

Table 2: A summary of the status of DukeMTMC and MS-Celeb-1M after their June 2019 retractions.

After the retraction, the authors of the DukeMTMC dataset issued an apology in *The Chronicle*, Duke's student newspaper, noting that the data collection had violated IRB guidelines in two respects: "Recording outdoors rather than indoors, and making the data available without protections." [71] However, this explanation did not appear on the website that hosted the dataset, which was simply taken down, meaning that not all users looking for the dataset would encounter this information. The retraction of MS-Celeb-1M fared worse: Microsoft never stated ethical motivations for removing the dataset, though the removal followed soon after multiple reports critiquing the dataset for privacy violations [36]. Rather, according to reporting by *The Financial Times*, Microsoft stated that the dataset was taken down "because the research challenge is over" [59]. The website that hosted MS-Celeb-1M is also no longer available. Neither retraction included calls to not use the data.

The disappearance of the websites also means that license information is no longer available through these sites. We were able to locate the DukeMTMC license through GitHub repositories of derivatives. We were unable to locate the MS-Celeb-1M license—which prohibits the redistribution of the dataset or derivatives—except through an archived version.[3] We discuss shortcomings of dataset licenses in Section 6.

We also identified public efforts to access and preserve these datasets, perhaps indicating confusion about the substantive meaning of the dataset's retractions. We found three and two Reddit posts inquiring about the availability of DukeMTMC and MS-Celeb-1M, respectively, following their retraction. Two of these posts (one for each dataset) noted or referred to investigations about potential privacy violations, but still inquired about where the dataset could be found.

In contrast to the retractions of DukeMTMC and MS-Celeb-1M, the retraction of TinyImages was more clear. On the dataset's website, the creators ask that "the community to refrain from using it in future and also delete any existing copies of the dataset that may have been downloaded" [1].

Tension with archival efforts. Outside of questions of efficacy, retraction can come into tension with efforts to archive datasets. Datasets are often seeded on the platforms Academic Torrents and Internet Archive. For example, while the creators TinyImages asked that existing copies of the dataset be deleted, it had already previously been added to both Academic Torrents and Internet Archive. The 2011 version of ImageNet, which contains offensive images that were later removed from official versions, had also previously been added to both sites. We found two Reddit posts emphasizing the need of data preservation following Tiny Images' removal. In work critiquing machine learning datasets, Crawford and Paglen [20] note the issue of "inaccessible or disappearing datasets," writing that

[3] An archived version from April 2017 (found via [36]) is available at `http://web.archive.org/web/20170430224804/http://msceleb.blob.core.windows.net/ms-celeb-v1-split/MSR_LA_Data_MSCeleb_IRC.pdf`.

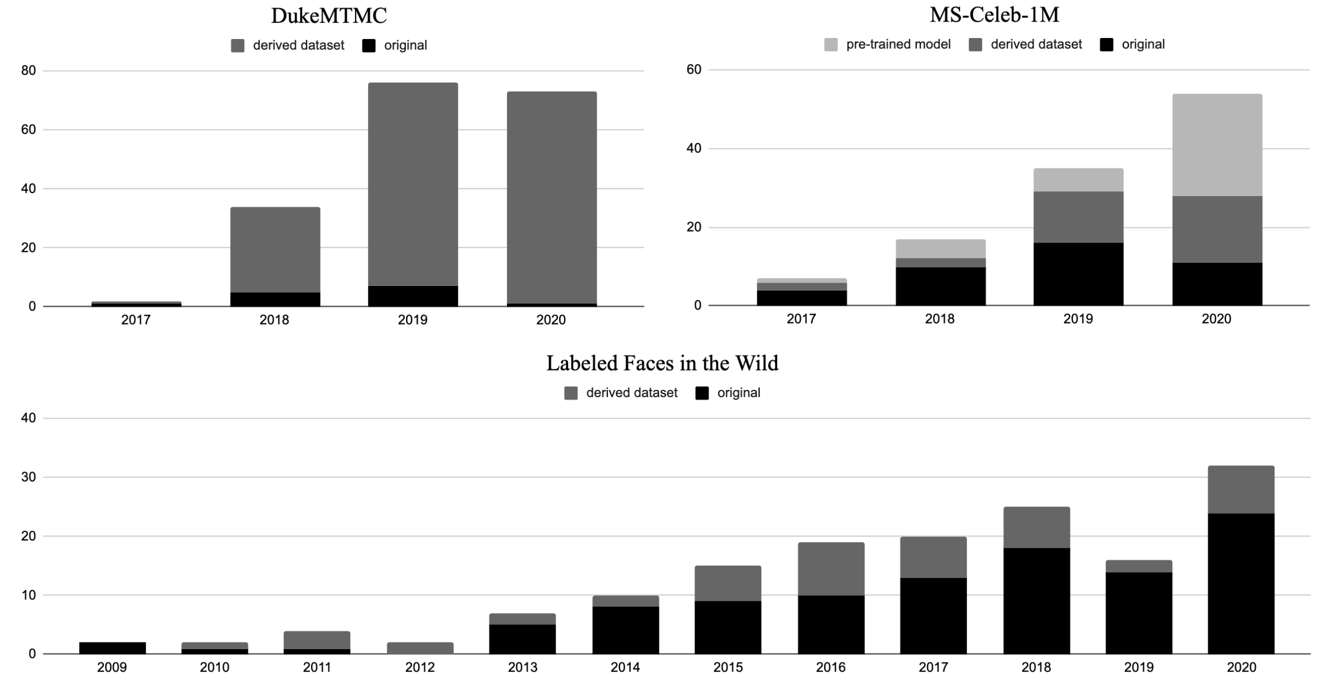

Figure 2: The use of DukeMTMC, MS-Celeb-1M, LFW, and their derivatives over time. DukeMTMC and MS-Celeb-1M were retracted in June 2019, but continued to be used in 2020—largely, through derivatives.

"If they are, or were, being used in systems that play a role in everyday life, it is important to be able to study and understand the worldview they normalize. Developing frameworks within which future researchers can access these data sets in ways that don't perpetuate harm is a topic for further work," illustrating this tension between regulation by retraction and archival interests.

4 Derivatives raise new ethical questions

Machine learning datasets often serve simultaneous roles as a specific tool (e.g., a benchmark on a particular task) and as a collection of raw material that can be otherwise leveraged. The latter role—whose impact may even eclipse the former—can be seen through a dataset's use in derivatives. Widespread derivative creation can be a success of resource-sharing in the machine learning community as it reduces the cost of obtaining data. However, it also means that a dataset's ethics are far from stagnant.

We identify four ways in which a derivative can raise ethical considerations, from which we categorized the 41 derivatives we identified (see Table 1). A derivative can alter the purpose of a dataset by enabling new applications or by training a publicly-available model, and can alter the composition of a dataset by adding annotations or by post-processing. When we say that a derivative raises new ethical considerations, we do not mean to imply that the creation of the derivative (or the parent dataset) is necessarily unethical.

New application. Either implicitly or explicitly, modifications of a dataset can enable new applications, which may warrant discussion. Twenty-one of 41 derivatives we identified fall under this category. We present several examples:

- DukeMTMC-ReID is a person re-identification benchmark, while the original DukeMTMC introduced a multi-target multi-camera tracking benchmark. While these problems are similar, they may have different motivating applications. We find that DukeMTMC-ReID is used much more frequently than DukeMTMC, implying that person re-identification has been the primary application of data collected through DukeMTMC. Several papers flagged by MegaPixels [36] using DukeMTMC data for ethically dubious purposes use DukeMTMC-ReID.

- Multiple derivatives of LFW label the original face images with attributes including race, gender, and attractiveness. The Racial Faces in the Wild dataset also groups images in MS-Celeb-1M by race. These labels enable new applications that may be ethically problematic, including the classification and retrieval of people via sensitive attributes.

- SMFRD is a derivative of LFW that adds face masks to its images. It is motivated by face recognition applications during the COVID-19 pandemic, when many people wear face-covering masks. "Masked face recognition" has been criticized for violating the privacy of those who may want to conceal their face (e.g., [55, 80]).

Pre-trained models. We found six model classes that were commonly trained on MS-Celeb-1M. Across these six classes, we found 21 GitHub repositories that released models pre-trained on MS-Celeb-1M. These pre-trained models can be used out-of-the-box to perform face recognition or can be used for transfer learning. Because the models can already compute salient feature representations of faces, they can be used as the basis for other tasks. This enables the use of MS-Celeb-1M for a wide range of applications, albeit in a more indirect way. Additionally, there remain questions about the effect of biases in training data on pre-trained models and their downstream applications (explored in [68]).

New annotations. The annotation of data can also result in privacy and representational harms. (See section 3.1 of [61] for a survey of work discussing representational concerns.) Seven of 41 derivatives fall under this category. Among the derivatives we examined, four annotated the data with gender, three with race or ethnicity, and two with additional traits including "attractiveness."

Other post-processing. Other derivatives neither repurpose the data for new applications nor contribute annotations. Rather, these derivatives are designed to aid the original task with more subtle modifications. Still, even minor modifications can raise ethical questions. For example, Datasheets for Datasets [28] includes a question about the potential effects of preprocessing or cleaning.[4]

- Five of 41 derivatives (each of MS-Celeb-1M) "clean" the original dataset, creating a more accurate set of images from the original, which is known to be noisy. This process often reduces the number of images significantly, after which, we may be interested in the resulting composition. Does the cleaning process reduce the number of images of people of a particular demographic group, for example? Such a shift may impact the downstream performance of such a dataset.
- Five of 41 derivatives (each of LFW) align, crop, or frontalize images in the original dataset. Here, too, we may ask about how such techniques perform on different demographic groups.

5 Technological and social change raise new ethical questions

We now turn to the question of how ethical considerations associated with a dataset change over time and may not be fully determined at the time of release. We do not seek to analyze the extent to which concerns can be anticipated at the time of release. Rather, we caution that—in practice—ethical concerns may arise or become apparent long after a dataset is created, raising implications for the timeline of effective interventions.

For this analysis, we do not use DukeMTMC and MS-Celeb-1M as they are relatively recent and thus less fit for longitudinal analysis. The two datasets we study are LFW, which was released in 2007, and ImageNet, which is widely considered to be among the most impactful datasets ever in machine learning and has also been at the center of ethical concerns involving datasets. ImageNet was released in 2009 and includes 14 million images annotated with 21,000 unique labels. In 2019, researchers revealed that many of the images in the "people" category of the dataset were labeled with misogynistic and racial slurs and perpetuated stereotypes, after which images in these categories were removed.

We focus on two broad ways that ethical considerations can change: through technological change and through social change.

[4] The question, in full, is: "Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?"

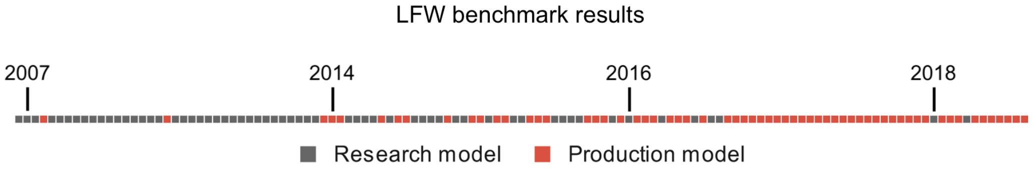

Figure 3: A visualization of the 123 benchmark results recorded on LFW's website according to whether they come from research papers or production models (as indicated on the website). The top performing models early on were almost all research models, whereas more recently, almost all are production models. We use this as a rough visualization of the rise of production facial recognition systems. The year labels mark the first model that was indicated to be from that year on the website.

5.1 Technological change shapes dataset ethics

Dataset use in production settings introduces and amplifies possible harm in comparison to research use. The improvement of technology facilitates this shift. We discuss how both LFW and ImageNet were introduced at a time when their problems of interest—face recognition and object classification, respectively—were relatively nascent. In the years since their introductions, both face recognition and object recognition have found widespread use in production settings, as have the datasets themselves. This transition—from research use to production use—is, in some sense, a sign of success of the dataset. Benchmark datasets in machine learning are typically introduced for problems that are not yet viable in production use cases; and should the benchmark be successful, it will help lead to the realization of real-world application.

LFW was introduced in 2007 to benchmark face verification. It is considered the first "in-the-wild" face recognition benchmark, designed to help face recognition improve in real-world, unconstrained settings (as compared to previous research, which had mostly been done to recognize faces captured in laboratory settings [63]). The production use of the dataset was unviable in its early years, one indication being that the benchmark performance on the dataset was poor.[5]

But over time, LFW became a standard benchmark and technology improved. This change can be seen through the benchmarking of commercial systems on the dataset (Figure 3). For example, of the 25 most recent results given on the LFW website for the "unrestricted, labeled outside data" setting, 23 were commercial systems. Of these, we found four commercial systems that advertised LFW accuracy on their website despite the creators of LFW noting that the dataset should not be used to conclude that an algorithm is suitable for production use. (We do not speculate on how else the commercial systems have been tested.) More broadly, the production use of facial recognition systems in applications such as surveillance or policing have caused backlash—especially because of disparate performance on minority groups. The rise of production facial recognition systems has raised new ethical concerns for LFW.

When ImageNet was introduced, object classification was still in its very early stages. Today, as real-world use of such technology has become widespread, ImageNet has become a common source for pre-training. Even as the dataset's terms of service specify non-commercial use, the dataset is commonly used in pre-trained models released under commercial licenses. (We discuss the production use of ImageNet in greater detail in Section 6.) Weights pre-trained on ImageNet come built-in in popular machine learning frameworks such as Keras and PyTorch. Here, too, the dataset has become used in production settings, raising new ethical concerns.

5.2 Social change shapes dataset ethics

We now turn to how social change can affect the ethics of a dataset. We take "social change" to mean several things. On one hand, we use the term in the sense of Birhane and Cummins when they write that "what society deems fair and ethical changes over time" [14]. We also use the term to describe growing awareness of ethical issues among researchers and practitioners, as well as shifting values and incentives within academia. Social change may also respond to technological change. We do not seek to disaggregate these various factors when describing the evolving ethics of datasets, nor do we try to discern cause and effect.

[5] Consider the benchmark task of face verification with unrestricted outside training data (the easiest of tasks proposed in [42]). The best reported performance by 2010 was 0.8445 ± 0.0046, achieved in [18].

Understanding of LFW's composition in terms of diversity has changed significantly over time. The 2007 technical report introducing LFW describes the dataset as exhibiting "natural variability in pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality" [42]. Yet a disclaimer added to the dataset's website in 2019 emphasizing how the dataset should not be used to verify commercial systems notes lacking variability among demographic groups, as well as in pose, lighting, occlusion, and resolution [3]. The earlier statement is meant to distinguish the dataset from its predecessors, which were relatively small and typically included images collected in laboratory settings [63]. We suggest several factors at play in the latter statement. In the last decade, standards for representation—both in the context of machine learning datasets and society as a whole—have changed. In relation to machine learning datasets, recent research and reporting has demonstrated disparate performances in machine learning algorithms, particularly in face recognition [16]. Research has highlighted how imbalanced datasets can lead to these disparities, and datasets such as Casual Conversations [38] and Diversity in Faces [52] seek to address this issue. (Here, we note that the Diversity in Faces dataset, in turn, faced backlash for being collected without consent [67], illustrating ongoing tensions between different ethical principles.)

We also consider recent ethical concerns involving ImageNet. We observe that work critiquing the dataset, most notably [20] and [62], first appeared nearly a decade after its release (even if issues were known to some earlier). As it is reasonable to assume that the labels used in ImageNet would have been considered offensive in 2009, the lag between the dataset's release and the removal of such labels is worth considering. We offer three hypotheses as to why there was such a lag. We hypothesize that growing public concern over machine learning applications, particularly in facial recognition, may have led to the publication of such work. Issues involving face datasets have received significant public attention—the article by Crawford and Paglen [20] accompanied several art exhibitions and the topic has been covered by many media outlets (e.g., [59, 54, 67]). Relatedly, we hypothesize that changing academic incentives led to the publication of this work. Related work highlighting assumptions underlying classification schemes [11, 44] have been published in FAccT, a machine learning conference focused on fairness, accountability, and transparency that was only founded in 2018. Finally, we hypothesize that norms involving the ethical responsibility of dataset creators and machine learning researchers more generally have shifted in recent years. These norms are still evolving; responses to recently-introduced ethics-related components of paper submission have been mixed [5].

(The labels in ImageNet are inherited from WordNet, a lexical database that was introduced in 1990. In turn, WordNet was constructed with help from Roget's thesaurus, which dates back as far as 1805 [57]. Thus, even as social norms progress, artifacts from older times and with different intended uses may be inherited.)

The life cycles of both LFW and ImageNet exhibit a lag between release and ethics-related inflection points. Such lags figure into the effectiveness of different interventions for harm mitigation.

6 Effectiveness of licenses

Licenses, or terms of use, are legal agreements between the creator and users of datasets, and often dictate how the dataset may be used, derived from, and distributed. There are many possible reasons for issuing licenses: respecting inherited licenses or copyright, maintaining exclusive commercial usage rights, reducing liability, ensuring proper attribution is received, etc. Here, we focus on the role of a license in harm mitigation, i.e., as a tool to restrict unintended and potentially harmful uses of a dataset.

6.1 Analysis of licenses

Through an analysis of licensing restrictions of DukeMTMC, MS-Celeb-1M, LFW, and their derivatives, we found shortcomings of licenses as a tool for mitigating harms.

Licenses do not effectively restrict production use. We analyzed the licensing information for DukeMTMC, MS-Celeb-1M, and LFW, and determined the implications for production use. Datasets are at a greater risk to do harm in production settings, where characteristics of a dataset directly affect people.

DukeMTMC is released under the CC BY-NC-SA 4.0 license, meaning that users may freely share and adapt the dataset, as long as attribution is given, it is not used for commercial purposes, derivatives are shared under the same license, and no additional restrictions are added to the license. Benjamin et al. [10] note many possible ambiguities in a "non-commercial" designation for a dataset. We emphasize, in particular, that this designation allows the possibility for non-commercial production use. Models deployed by nonprofits and governments maintain risks associated with commercial models.

MS-Celeb-1M is released under a Microsoft Research license agreement,[6] which has several specifications, including that users may "use and modify this Corpus for the limited purpose of conducting non-commercial research." The legality of using models trained or pre-trained on this data remains unclear—a recurring theme throughout the remainder of this section.

LFW was released without any license. In 2019, a disclaimer was added on the dataset's website, indicating that the dataset "should not be used to conclude that an algorithm is suitable for any commercial purpose." [3] The lack of an original license meant that the dataset's use was entirely unrestricted until 2019. Furthermore, while it includes useful guiding information, the disclaimer does not hold legal weight. Additionally, through an analysis of results given on the LFW website [3], we found four commercial systems that clearly advertised their performance on the datasets, though we do not know if the disclaimer is intended to discourage this behavior. Because LFW is a relatively small dataset, its use as training data in production settings is unlikely. Risk remains, as the use of its performance as a benchmark on commercial systems can lead to overconfidence, both among the system creators and potential clients. Raji et al. [64] highlight various ethical considerations when auditing facial recognition systems.

Derivatives do not always inherit original terms. Both DukeMTMC and MS-Celeb-1M, according to their licenses, may only be used for non-commercial use. (This analysis does not apply to LFW, which was released with no license.) We analyzed available derivatives of each dataset to see if they include a non-commercial use designation. All four DukeMTMC derivative datasets included the designation. Four of seven MS-Celeb-1M derivative datasets included the designation. Only three of 21 repositories containing models pre-trained on MS-Celeb-1M included the designation.

Thus, we find mixed results of license inheritance. We note that DukeMTMC's license specifies that derivatives must include the original license. Meanwhile, MS-Celeb-1M's license, which prohibits derivative distribution in the first place, is no longer publicly accessible, perhaps partially explaining the results. Licenses are only effective if actively followed and inherited by derivatives.

The loose licenses associated with the pre-trained models are particularly notable. Of the 21 repositories containing pre-trained models, seven contained the MIT license, one contained the Apache 2.0 license, one contained the BSD-2-Clause license, and nine contained no license at all.

6.2 Commercial use of models trained on non-commercial data

In this section, we seek to understand whether models trained on datasets released for non-commercial research are being used commercially. Whether or not such use is legal, it can exacerbate the real-world harm caused by datasets.

Due to the obvious difficulties involved in studying this question, we approach it by studying online discussions. We identified 14 unique posts on common discussion sites that inquired about the legality of using pre-trained models that were trained on non-commercial datasets.[7] These 14 posts yielded numerous responses representing a wide range of suggestions. The question of legality is not one we seek to answer; for our purposes, it is merely a window into commercial practices.

Evidence of commercial use. From these posts, we found anecdotal evidence that non-commercial dataset licenses are sometimes ignored in practice. One response reads: "More or less everyone (individuals, companies, etc) operates under the assumption that licences on the use of data do not apply to models trained on that data, because it would be extremely inconvenient if they did." Another response reads: "I don't know how legal it really is, but I'm pretty sure that a lot of people develop algorithms that are based on a pretraining on ImageNet and release/sell the models without caring about legal issues. It's not that easy to prove that a production model has been pretrained on ImageNet ..."

Commonly-used computer vision frameworks such as Keras and PyTorch include models pre-trained on ImageNet, making the barrier for commercial use low.

[6] The license is no longer publicly available. An archived version is available here: `http://web.archive.org/web/20170430224804/http://msceleb.blob.core.windows.net/ms-celeb-v1-split/MSR_LA_Data_MSCeleb_IRC.pdf`

[7] We identified these posts via four Google searches with the query "pre-trained model commercial use." We then searched the same query on Google with "site:www.reddit.com," "site:www.github.com," "site:www.twitter.com," and "site:www.stackoverflow.com." These are four sites where questions about machine learning are posted. For each search, we examined the top 10 sites presented by Google. Within relevant posts, we also extracted any additional relevant links included in the discussion.

Responses from model distributors. These posts resulted in responses from representatives of three pre-trained model creators: representative of the popular Keras and PyTorch frameworks, as well as the creator of LIP_JPPNet, a pre-trained model for "human parsing and pose estimation."[8]

The former two suggested that such use is generally allowed, but that they could not provide an official answer. A representative for Keras wrote that "In the general case, pre-trained weight checkpoints have their own license which isn't inherited from the license of the dataset they were trained on. This is not legal advice, and you should consult with a lawyer." A representative for PyTorch wrote that according to their legal team's guidance, "weights of a model trained on that data may be considered derivative enough to be ok for commercial use. Again, this is a subjective matter of comfort. There is no publishable 'answer' we can give."

A representative of LIP_JPPNet suggested that the user's concern was correct, and that "You can train the code on your own datasets to get a model for commercial use."

7 Dataset management and citation

Building on our previous findings, we turn to the role of dataset management and citation in harm mitigation. By dataset management, we mean storing a dataset and information about it. By dataset citation, we mean the referencing of a dataset used in research, with the particular aim of facilitating access to the dataset itself and supporting information. Together, management and citation can support three processes that facilitate harm mitigation: transparency, documentation, and tracking. Though our focus is on harm mitigation, dataset management and citation also help scientific communities function more efficiently [75].

Our main finding in this section is that the existing norms in the machine learning community for both dataset management and citation fall short of facilitating transparency, documentation, and tracking. Throughout this section, our discussions of datasets also apply to pre-trained models.

7.1 Dataset management and citation can enable harm mitigation

We give three reasons for why dataset management and citation are important for mitigating harms caused by datasets.

Documentation. Access to dataset documentation facilitates responsible dataset use. Documentation can provide information about a dataset's composition, its intended use, and any restrictions on its use (through licensing information, for example). Many researchers have proposed documentation tools for machine learning datasets with harm mitigation in mind [28, 9]. Dataset management and citation can ensure that documentation is easily accessible. Proper management ensures that documentation of a dataset is available even if the dataset itself is not or is no longer publicly accessible. In Section 3 and Section 6, we discussed how retracted datasets no longer included key information such as licensing information, potentially leading to confusion. For example, with MS-Celeb-1M's license no longer publicly available, the license status of derivative datasets, pre-trained models, and remaining copies of the original is unclear.

Transparency (and accountability). Dataset citation facilitates transparency in dataset use, in turn allowing for accountability. By clearly indicating the dataset used and where information about the dataset can be found, researchers become accountable for ensuring the quality of the data and its proper use. Different stakeholders, such as the dataset creator, program committees, and other actors can then hold researchers accountable. For example, if proper citation practices are followed, peer reviewers can more easily check whether researchers are in compliance with the terms of use of the datasets they have used.

Tracking. Large-scale analysis of dataset use—as we do in this paper—can illuminate a dataset's impact and potential avenues of risk or misuse. This knowledge can allow dataset creators to update documentation, better establishing intended use. Citation infrastructure supports this task by collecting such use in an organized manner. This includes both tracking the direct use of a dataset in academic research, as well as the creation of derivatives.

[8] https://github.com/Engineering-Course/LIP_JPPNet

Goal	Shortcomings of current management and citation practices
Documentation	– Documentation is not easily accessible from citations – Documentation is not persistent (i.e., it may not remain available over time)
Transparency	– Dataset use is not clearly specified in academic papers – Documentation is not easily accessible from citations
Tracking	– Datasets do not have trackable identifiers (such as DOIs) – Associated papers are not robust proxies – Derivatives are not trackable

Table 3: Summary of how current dataset management and citation practices fail to support harm mitigation.

Reference	Attempted disambiguation
"Experiments were performed on four of the largest ReID benchmarks, i.e., Market1501 [45], CUHK03 [17], DukeMTMC [33], and MSMT17 [40] . . . DukeMTMC provides 16,522 bounding boxes of 702 identities for training and 17,661 for testing."	Here, the dataset is called DukeMTMC and the citation [33] is of DukeMTMC's associated paper. However, the dataset is described as an ReID benchmark. Moreover, the statistics given exactly match the popular DukeMTMC-ReID derivative (an ReID benchmark). This leads us to believe DukeMTMC-ReID was used.
"We used the publicly available database Labeled Faces in the Wild (LFW)[6] for the task. The LFW database provides aligned face images with ground truth including age, gender, and ethnicity labels."	The name and reference both point to the original LFW dataset However, the dataset is described to contain aligned images with age, gender, and ethnicity labels. The original dataset contains neither aligned images nor any of these annotations. There are, however, many derivatives with aligned versions or annotations by age, gender, and ethnicity. Since no other description was given, we were unable to disambiguate.
"MS-Celeb-1M includes 1M images of 100K subjects. Since it contains many labeling noise, we use a cleaned version of MS-Celeb-1M [16]."	The paper uses a "cleaned version of MS-Celeb-1M," but the particular one is not specified. (There are many cleaned versions of the dataset.) The citation [16] is to the original MS-Celeb-1M's associated paper and no further description is given. Therefore, we were unable to disambiguate.

Table 4: Examples of dataset references that were challenging to disambiguate.

7.2 Current practices fall short

Above, we established key functions of dataset management and citation in harm mitigation. Through our analysis of dataset use in machine learning research, we found that current practices fall short in achieving these functions. Our findings are summarized in Table 3.

Dataset management practices raise concerns for persistence. Whereas other fields utilize shared repositories, machine learning datasets are often managed through the sites of individual researchers or academic groups. None of the 38 datasets in our analysis are managed through such repositories. Unsurprisingly, we found that some datasets were no longer maintained (which is different from being retracted). We were only able to find information about D31 and D38 through archived versions of sites found via the WayBack machine. And even after examining archived sites, we were unable to locate information about D6, D17, and D25. Another consequence is the lack of persistence of documentation. Ideally, information about a dataset should remain available even if the dataset itself is no longer available. But we found that after DukeMTMC and MS-Celeb-1M were taken down, so too were the sites that contained their terms of use.

Dataset references can be difficult to disambiguate. Clear dataset citation is important for harm mitigation for transparency and documentation. However, datasets are not typically designated as independent citable research

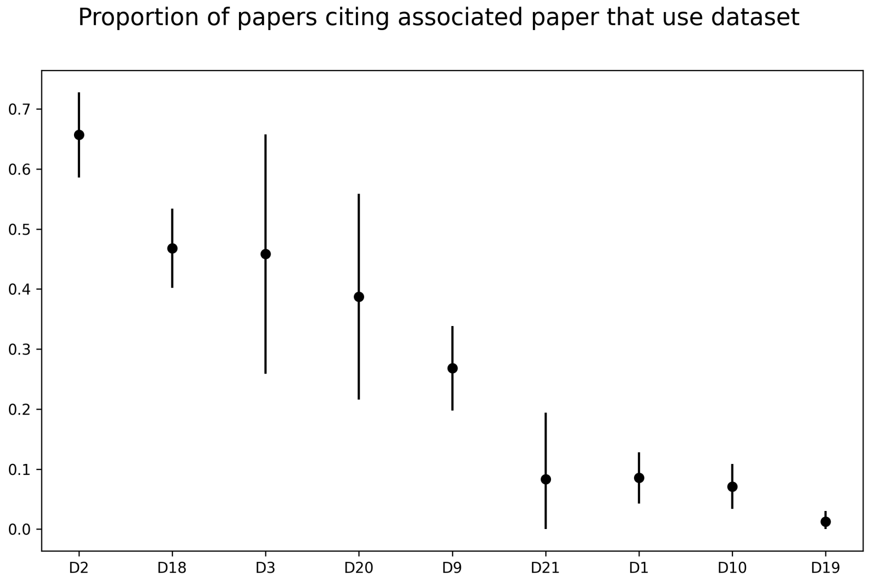

Figure 4: Papers citing associated papers often do not use the associated dataset. The proportion that do varies greatly across different datasets. Here, we include associated papers for which we sampled at least 20 citing papers, and show 95 percent confidence intervals.

objects like academic papers are. This is evidenced by a lack of standardized permanent identifiers, such as DOIs. None of the 38 datasets we encountered in our analysis had such identifiers. (Datasets are often assigned DOIs when added to shared data repositories.)

Without dataset-specific identifiers, we found that datasets were typically cited with a combination of the dataset's name, a description, and paper citations. In many cases, an associated paper is cited—a paper through which a dataset was introduced or that the dataset's creators request be cited. In some cases, a dataset does not have a clear associated paper. For example, D31 was not introduced in an academic paper and D20's creators suggest three distinct academic papers that may be cited. This practice can lead to challenges in identifying and accessing the dataset(s) used in a paper, especially when the name, description, or citation conflict.

In our analysis, 42 papers included dataset references that we were unable to fully disambiguate. Oftentimes, this was a result of conflating a dataset with its derivatives. For example, we found nine papers that suggested that images in LFW were annotated with attributes or keypoints, but did not specify where these annotations were obtained. (LFW only contains images labeled with identities and many derivatives of LFW include annotations.) Similarly, seven papers indicated that they used a cleaned version of MS-Celeb-1M, but did not identify the particular derivative. We were able to disambiguate the references in 404 papers using a dataset or a derivative, but in many of these instances, making a determination was not direct (for instance, see the first example in Table 4).

Datasets and documentations are not directly accessible from citations. We find that accessing datasets from papers is not currently straightforward. While data access requirements, such as sections designated to specifying where datasets and other supporting materials may be found, are common in other fields, they are rare in machine learning. We sampled 60 papers from our sample that used DukeMTMC, MS-Celeb-1M, LFW, or one of their derivative datasets, and only six provided access information (each as a URL).

Furthermore, the descriptors we mentioned above—paper citations, name, and description—do not offer a direct path to the dataset. The name of a dataset can sometimes be used to locate the dataset via web search, but this works poorly in many instances—for example, when a dataset is not always associated with a particular name or when the dataset is not even available. Datasets D27, D28, D31, D32, and D38 are not named. Citations of an associated paper also do not directly convey access information. As an alternate strategy, we were able to locate some datasets by searching for personal websites of the dataset creators or of their associated academic groups. However, as we mentioned earlier, we were still unable to locate D6, D17, and D25, even after looking at archived versions of sites.

Current infrastructure makes tracking dataset use difficult. A lack of dataset citations also makes it difficult to track dataset use. Citations of associated papers are not necessarily effective proxies in this respect. On one hand, the proportion of papers citing an associated paper that use the corresponding dataset varies significantly (see Figure 4). This is because papers citing an associated paper may be referencing other ideas mentioned by the paper. On the other hand, some datasets may be commonly used in papers that do not cite a particular associated paper. Of the papers we found to use DukeMTMC-ReID, 29 cited the original dataset, 63 cited the derivative dataset, and 50 cited both. Furthermore, some datasets may not have a clear associated paper, and various implementations of pre-trained models are unlikely to have associated papers. Thus, associated papers—as currently used—are an exceedingly clumsy way to track the use of datasets.

Tracking derivative creation presents an even greater challenge. Currently, there is no clear way to identify derivatives. The websites of LFW and DukeMTMC (the latter no longer online), maintained lists of derivatives. However, our analysis reveals that these lists are far from complete. Proponents of dataset citation have suggested the inclusion of metadata indicating provenance in a structured way (thus linking a dataset to its derivatives) [30], but such a measure has not been adopted by the machine learning community.

8 Recommendations

In the last few years, there have been numerous recommendations for mitigating the harms associated with machine learning datasets, including better documentation [28, 9, 39], "interventionist" data collection approaches modeled on archives [43], calls for requiring informed consent from data subjects [62], and strategies for bias mitigation and privacy preservation [8, 73, 62, 81]. Our own recommendations draw from and build on this body of work, and aren't meant to replace existing ideas for harm mitigation.

That said, previous harm-mitigation approaches primarily consider dataset creation. As we have shown above, ethical impacts are hard to anticipate and address at dataset creation time and thus we argue that harm mitigation requires stewarding throughout the life cycle of a dataset. Our recommendations reflect this understanding.

We contend that the problem cannot be left to any one stakeholder such as dataset creators or IRBs. We propose a more distributed approach in which many stakeholders share responsibility for ensuring the ethical use of datasets. Our approach assumes the willingness and cooperation of dataset creators, program committees, and the broader research community; addressing harms from callous or malicious users or outside of the research context is outside the scope of our recommendations.

8.1 Dataset creators

We make two main recommendations for dataset creators, both based on the normative influence they can exert and based on the harder constraints they can impose on how datasets are used.

Make ethically-salient information clear and accessible. Dataset creators can place restrictions on dataset use through licenses and provide other ethically-salient information through other documentation. But in order for these restrictions to be effective, they must be clear. In our analysis, we found that licenses are often insufficiently specific. For example, when restricting the use of a dataset to "non-commercial research" creators should be explicit about whether this also applies for models trained on the dataset. It may also be helpful to explicitly prohibit specific ethically questionable uses. The Casual Conversations dataset does this ("Participants will not ... use the Casual Conversations Dataset to measure, detect, predict, or otherwise label the race, ethnicity, age, or gender of individuals, label facial attributes of individuals, or otherwise perform biometric processing unrelated to the Purpose") [38]. The Montreal Data License [10] is an example of a license that allows for more specific designations. Going beyond licenses, Datasheets for Datasets provides a standard for detailed and clear documentation of ethically-salient information [28].

In order for licenses or documentation to be effective, they need to be accessible. Licenses and documentation should be persistent, which can be facilitated through the use of standard data repositories. Dataset creators should also set requirements for dataset users and creators of derivatives to ensure that this information is easy to find from citing papers and derived datasets.

Actively steward the dataset and exert control over use. Throughout our analysis, we show how ethical considerations can evolve over time. Dataset creators should continuously steward a dataset, actively examining how it may be misused, and making updates to license, documentation, or access restrictions as necessary. A minimal access restriction is for users to agree to terms of use. A more heavyweight process in which dataset creators make case-by-case decisions about access requests can be used in cases of greater risk. The Fragile Families Challenge is an example of this [51].

Based on our analysis in Section 3 and Section 4, derivative creation often introduces new ethical risks. Thus, we suggest that dataset creators use procedural mechanisms to control derivative creation, for example, by requiring explicit permission be obtained to create a derivative.

We recognize that dataset stewarding increases the burden on dataset creators. In our discussions with dataset creators, we heard that creating datasets is already an undervalued activity and that a norm of dataset stewardship might further erode the incentives for creators. We acknowledge this concern, but maintain that there is an inherent tension between ethical responsibility and minimizing burdens on dataset creators. One solution is for dataset creation to be better rewarded in the research community; some of our suggestions for program committees below may have this effect.

8.2 Conference program committees

In the machine learning community and the broader computer science community, the primary publication venues are the peer-reviewed proceedings of conferences, and the prestige associated with these publications is the core of the reward structure for researchers. Thus, conference program committees (PCs) hold immense power to set ethical standards for the research community including both dataset creators and users. Of course, PCs have little power over commercial use of datasets that is not motivated by publications. Ethics review as part of peer review is a recent development at machine learning conferences but has a longer history in the computer security community.

Use ethics review to encourage responsible dataset use. PCs are in a position to govern both the creation of datasets (and derivatives) and the use of datasets through ethics reviews of the associated papers. PCs should develop clear guidelines for ethical review. For example, PCs can require researchers to clearly state the datasets used, justify the reasons for using those datasets, and to certify that they complied with the terms of use of each dataset. Some conferences, such as NeurIPS, already have ethics guidelines relating to dataset use.

Encourage standard dataset management and citation practices. PCs should consider imposing standardized dataset management and citation requirements. This can include requiring dataset creators to upload their dataset and supporting documentation to a public data repository. Detailed guidelines on effective dataset management and citation practices can be found in [75]. The role of PCs is particularly important for dataset management and citation, as these practices benefit from community-wide adoption of standards.

Introduce a dataset-specific track. NeurIPS now includes a track specifically for datasets. The introduction of such tracks facilitates more careful and tailored ethics reviews for datasets. The journal Scientific Data is devoted entirely to describing datasets.

Allow advance review of datasets and publications. We tentatively recommend that conferences can allow researchers to submit proposals for datasets prior to creation. By receiving preliminary feedback, researchers can be more confident that their dataset both will be valued and will pass initial ethics reviews. This mirrors the concept of "registered reports" in which a proposed study is peer reviewed before it is undertaken and provisionally accepted for publication *before* the outcomes are known, as a way to counter publication biases.

8.3 Institutional Review Boards

Historically, Institutional Review Boards (IRBs) have played a fundamental role in regulating research ethics, and researchers have recently called for greater IRB oversight in dataset creation [62]. IRBs have certain natural advantages in regulating datasets. IRBs may have more ethics expertise than program committees; IRBs are also able to review datasets prior to their creation. Thus, IRBs can prevent harms that occur during the creation process.

However, conceived first to address biomedical research, IRBs have been an imperfect fit for data-centered research. Notably "human subjects research" has a narrow definition and thus many of the datasets (and associated research) that have caused ethical concern in machine learning would not fall under the purview of IRBs. An even more significant limitation is that IRBs are not allowed to consider downstream harms [53][9].

Unless and until the situation changes, our primary recommendation regarding IRBs is for researchers to recognize that research being approved by the IRB does not mean that it is "ethical," and for IRBs themselves to make this as clear as possible.

8.4 Dataset users and other researchers

At a minimum, dataset users should comply with the terms of use of datasets. But their responsibility goes beyond compliance. They should also carefully study the accompanying documentation and analyze the appropriateness of using the dataset for their particular application (e.g., whether dataset biases may propagate to models). Dataset users should also clearly indicate what dataset is being used in their research papers and ensure that readers can access the dataset based on the information provided. As we showed in Section 7, traditional paper citations often lead to ambiguity.

Our findings show how a dataset's impact is not fully understood at the time of its creation. We recommend that the community systematize the retrospective study of datasets to understand their ethical implications, potential shortcomings, and misuse. Researchers should not wait until the problems become serious and there is an outcry.

It is especially important to understand how datasets and pre-trained models are being used in production settings, which our work does not address. Policy makers should consider legal requirements that encourage more transparency around the specifics of training datasets used in commercially deployed models.

9 Conclusion

The machine learning community is responding to a wide range of ethical concerns regarding datasets and asking fundamental questions about the role of datasets in machine learning research. We provide a new perspective. Through our analysis of the life cycles of three datasets, we showed how developments that occur after dataset creation can impact the ethical consequences, making them hard to anticipate a priori. We advocate for an approach to harm mitigation in which responsibility is distributed among stakeholders and continues throughout the life cycle of a dataset.

10 Acknowledgments

This work is supported by the National Science Foundation under Awards IIS-1763642 and CHS-1704444. We thank participants of the Responsible Computer Vision (RCV) workshop at CVPR 2021, the Princeton Bias-in-AI reading group, and the Princeton CITP Work-in-Progress reading group for useful discussion. We also thank Solon Barocas, Marshini Chetty, Mihir Kshirsagar, and Olga Russakovsky for their helpful feedback.

References

[1] 80 Million Tiny Images. `https://groups.csail.mit.edu/vision/TinyImages`.

[2] Gender Classification. `http://web.archive.org/web/20161026012553/http://fipa.cs.kit.edu/431.php`.

[3] Labeled Faces in the Wild Home. `http://vis-www.cs.umass.edu/lfw`.

[4] Trillionpairs. `http://trillionpairs.deepglint.com/overview`.

[5] Grace Abuhamad and Claudel Rheault. Like a Researcher Stating Broader Impact For the Very First Time. *Navigating the Broader Impacts of AI Research Workshop, NeurIPS*, 2020.

[6] Ankan Bansal, Rajeev Ranjan, Carlos D. Castillo, and Rama Chellappa. Deep Features for Recognizing Disguised Faces in the Wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 10–106, 2018.

[9] "The IRB should not consider possible long-range effects of applying knowledge gained in the research (e.g., the possible effects of the research on public policy) as among those research risks that fall within the purview of its responsibility." (45 CFR §46.111)

[7] Brian C. Becker and Enrique G. Ortiz. Evaluating Open-Universe Face Identification on the Web. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 904–911, 2013.

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *CoRR*, abs/1810.01943, 2018.

[9] 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.

[10] Misha Benjamin, P. Gagnon, N. Rostamzadeh, C. Pal, Yoshua Bengio, and Alex Shee. Towards Standardization of Data Licenses: The Montreal Data License. *ArXiv*, abs/1903.12262, 2019.

[11] Sebastian Benthall and Bruce D. Haynes. Racial Categories in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 289–298, New York, NY, USA, 2019. Association for Computing Machinery.

[12] Tamara Berg, Alexander Berg, Jaety Edwards, and David Forsyth. Who's In the Picture. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.

[13] L. Best-Rowden, Shiwani Bisht, Joshua C. Klontz, and Anil K. Jain. Unconstrained face recognition: Establishing baseline human performance via crowdsourcing. *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014.

[14] Abeba Birhane and Fred Cummins. Algorithmic Injustices: Towards a Relational Ethics. *ArXiv*, abs/1912.07376, 2019.

[15] Su Lin Blodgett and Brendan T. O'Connor. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *ArXiv*, abs/1707.00061, 2017.

[16] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[17] Aylin Caliskan, Joanna Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356:183–186, 2017.

[18] Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. Face recognition with learning-based descriptor. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2707–2714, 2010.

[19] Joseph Paul Cohen and Henry Z. Lo. Academic Torrents: A Community-Maintained Distributed Repository. In *Annual Conference of the Extreme Science and Engineering Discovery Environment*, 2014.

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

[21] Matthias Dantone, Juergen Gall, G. Fanelli, and L. Gool. Real-time facial feature detection using conditional regression forests. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.

[22] Jiankang Deng, J. Guo, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[23] Jiankang Deng, J. Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight Face Recognition Challenge. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2638–2646, 2019.

[24] Jiankang Deng, Yuxiang Zhou, and S. Zafeiriou. Marginal Loss for Deep Face Recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2006–2014, 2017.

[25] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets, 2020.

[26] Claudio Ferrari, Giuseppe Lisanti, Stefano Berretti, and Alberto Del Bimbo. Effective 3D based frontalization for unconstrained face recognition. *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1047–1052, 2016.

[27] Allen Institute for AI. Semantic Scholar API. `https://api.semanticscholar.org`.

[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. *CoRR*, abs/1803.09010, 2018.

[29] Mengran Gou, S. Karanam, WenQian Liu, O. Camps, and R. Radke. DukeMTMC4ReID: A Large-Scale Multi-camera Person Re-identification Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1425–1434, 2017.

[30] Paul T. Groth, H. Cousijn, Tim Clark, and C. Goble. FAIR Data Reuse – the Path through Data Citation. *Data Intelligence*, 2:78–86, 2020.

[31] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. *2009 IEEE 12th International Conference on Computer Vision*, pages 498–505, 2009.

[32] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 87–102, Cham, 2016. Springer International Publishing.

[33] Hu Han, Anil K. Jain, F. Wang, S. Shan, and Xilin Chen. Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2597–2609, 2018.

[34] Margot Hanley, Apoorv Khandelwal, Hadar Averbuch-Elor, Noah Snavely, and Helen Nissenbaum. An Ethical Highlighter for People-Centric Dataset Creation. *arXiv preprint arXiv:2011.13583*, 2020.

[35] A. Hanna, Emily L. Denton, A. Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.

[36] Adam Harvey and Jules LaPlace. Exposing.ai. `https://exposing.ai`, 2021.

[37] Tal Hassner, Shai Harel, E. Paz, and Roee Enbar. Effective face frontalization in unconstrained images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4295–4304, 2015.

[38] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and C. Canton-Ferrer. Towards measuring fairness in AI: the Casual Conversations dataset. *ArXiv*, abs/2104.02821, 2021.

[39] Sarah Holland, A. Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *ArXiv*, abs/1805.03677, 2018.

[40] Gary B. Huang, Vidit Jain, and Erik Learned-Miller. Unsupervised Joint Alignment of Complex Images. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[41] Gary B. Huang, Marwan A. Mattar, Honglak Lee, and Erik Learned-Miller. Learning to Align from Scratch. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 764–772, Red Hook, NY, USA, 2012. Curran Associates Inc.

[42] 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.

[43] 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*, FAT* '20, page 306–316, New York, NY, USA, 2020. Association for Computing Machinery.

[44] Zaid Khan and Yun Fu. One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

[45] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Attribute and simile classifiers for face verification. *2009 IEEE 12th International Conference on Computer Vision*, pages 365–372, 2009.

[46] Minxian Li, Xiatian Zhu, and S. Gong. Unsupervised Tracklet Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:1770–1782, 2020.

[47] Lin Liang, Rong Xiao, Fang Wen, and Jian Sun. Face Alignment Via Component-Based Discriminative Search. In *European Conference on Computer Vision*, pages 72–85. Springer, 2008.

[48] Shengcai Liao, Zhen Lei, Dong Yi, and S. Li. A benchmark study of large-scale unconstrained face recognition. *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014.

[49] Yutian Lin, L. Zheng, Zhedong Zheng, Yu Wu, and Y. Yang. Improving Person Re-identification by Attribute and Identity Learning. *ArXiv*, abs/1703.07220, 2019.

[50] Z. Liu, Ping Luo, Xiaogang Wang, and X. Tang. Deep Learning Face Attributes in the Wild. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015.

[51] Ian Lundberg, A. Narayanan, Karen Levy, and Matthew J. Salganik. Privacy, Ethics, and Data Access: A Case Study of the Fragile Families Challenge. *Socius*, 5, 2018.

[52] Michele Merler, N. Ratha, R. Feris, and John R. Smith. Diversity in Faces. *ArXiv*, abs/1901.10436, 2019.

[53] Jacob Metcalf. "The study has been approved by the IRB": Gayface AI, research hype and the pervasive data ethics gap. *Pervade Team*, Nov 2017.

[54] Cade Metz. Facial Recognition Tech Is Growing Stronger, Thanks to Your Face. `https://www.nytimes.com/2019/07/13/technology/databases-faces-facial-recognition-technology.html`, Jul 2019.

[55] Rachel Metz. Think your mask makes you invisible to facial recognition? Not so fast, AI companies say. `https://www.cnn.com/2020/08/12/tech/face-recognition-masks/index.html`, Aug 2020.

[56] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-Guided Feature Alignment for Occluded Person Re-Identification. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 542–551, 2019.

[57] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*, 3:235–244, 1990.

[58] Margaret Mitchell, Simone Wu, Andrew Zaldivar, P. Barnes, Lucy Vasserman, B. Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.

[59] Madhumita Murgia. Microsoft quietly deletes largest public face recognition data set. `https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2`, Jun 2019.

[60] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[61] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*, 2020.

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

[63] Inioluwa Deborah Raji and Genevieve Fried. About Face: A Survey of Facial Recognition Evaluation, 2021.

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

[65] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. *CoRR*, abs/1609.01775, 2016.

[66] Conrad Sanderson and Brian C. Lovell. Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference. In Massimo Tistarelli and Mark S. Nixon, editors, *Advances in Biometrics*, pages 199–208, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[67] Olivia Solon. Facial recognition's 'dirty little secret': Millions of online photos scraped without consent. `https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921`, Mar 2019.

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

[69] R. Stewart, M. Andriluka, and A. Ng. End-to-End People Detection in Crowded Scenes. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2325–2333, 2016.

[70] Y. Sun, Xiaogang Wang, and X. Tang. Deep Convolutional Network Cascade for Facial Point Detection. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[71] Carlo Tomasi. Letter: Video analysis research at Duke. `https://www.dukechronicle.com/article/2019/06/duke-university-video-analysis-research-at-duke-carlo-tomasi`, Jun 2019.

[72] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[73] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. In *Computer Vision – ECCV 2020*, pages 733–751, Cham, 2020. Springer International Publishing.

[74] M. Wang, W. Deng, Jiani Hu, Xunqiang Tao, and Y. Huang. Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2019.

[75] M. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. Boiten, Luiz Olavo Bonino da Silva Santos, P. Bourne, J. Bouwman, A. Brookes, Tim Clark, M. Crosas, I. Dillo, Olivier Dumon, S. Edmunds, C. Evelo, R. Finkers, Alejandra N. González-Beltrán, A. Gray, Paul Groth, C. Goble, J. Grethe, J. Heringa, P. 't Hoen, R. Hooft, Tobias Kuhn, Ruben G. Kok, J. Kok, S. Lusher, M. Martone, Albert Mons, A. Packer, Bengt Persson, P. Rocca-Serra, M. Roos, René C van Schaik, Susanna-Assunta Sansone, E. Schultes, T. Sengstag, Ted Slater, George O. Strawn, M. Swertz, Mark Thompson, J. van der Lei, E. V. van Mulligen, Jan Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

[76] Lior Wolf, Tal Hassner, and Yaniv Taigman. Effective Unconstrained Face Recognition by Combining Multiple Descriptors and Learned Background Statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1978–1990, 2011.

[77] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018.

[78] Yuehua Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Y. Yang. Exploit the Unknown Gradually: One-Shot Video-Based Person Re-identification by Stepwise Learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[79] Hao Xiao, Weiyao Lin, Bin Sheng, K. Lu, Junchi Yan, Jingdong Wang, Errui Ding, Yihao Zhang, and H. Xiong. Group Re-Identification: Leveraging and Integrating Multi-Grain Information. *Proceedings of the 26th ACM international conference on Multimedia*, 2018.

[80] Wudan Yan. Face-mask recognition has arrived-for better or worse. https://www.nationalgeographic.com/science/article/face-mask-recognition-has-arrived-for-coronavirus-better-or-worse-cvd, May 2021.

[81] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. *ArXiv*, abs/2103.06191, 2021.

[82] Zhong yuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Q. Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, Heling Chen, Yu Miao, Z. Huang, and Jinbi Liang. Masked Face Recognition Dataset and Application. *ArXiv*, abs/2003.09093, 2020.

[83] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial Landmark Detection by Deep Multi-task Learning. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 94–108, Cham, 2014. Springer International Publishing.

[84] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in vitro. *CoRR*, abs/1701.07717, 2017.

[85] Yuxiang Zhou and S. Zafeiriou. Deformable Models of Ears in-the-Wild for Alignment and Recognition. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 626–633, 2017.