Imperfect ImaGANation: Implications of GANs Exacerbating Biases on Facial Data Augmentation and Snapchat Selfie Lenses

Niharika Jain[1], **Alberto Olmo**[1], **Sailik Sengupta**[1]
Lydia Manikonda[2], **Subbarao Kambhampati**[1]

[1] Arizona State University
[2] Rensselaer Polytechnic Institute
{njain30, aolmoher, sailiks, rao}@asu.edu, manikl@rpi.edu

Abstract

Recently, the use of synthetic data generated by GANs has become a popular method to do data augmentation for many applications. While practitioners celebrate this as an economical way to obtain synthetic data for training data-hungry machine learning models, it is not clear that they recognize the perils of such an augmentation technique when applied to an already-biased dataset. Although one expects GANs to replicate the distribution of the original data, in real-world settings with limited data and finite network capacity, GANs suffer from mode collapse. Especially when this data is coming from online social media platforms or the web which are never balanced. In this paper, we show that in settings where data exhibits bias along some axes (eg. gender, race), failure modes of Generative Adversarial Networks (GANs) exacerbate the biases in the generated data. More often than not, this bias is unavoidable; we empirically demonstrate that given input of a dataset of headshots of engineering faculty collected from 47 online university directory webpages in the United States is biased toward white males, a state-of-the-art (unconditional variant of) GAN "imagines" faces of synthetic engineering professors that have masculine facial features and white skin color (inferred using human studies and a state-of-the-art gender recognition system). We also conduct a preliminary case study to highlight how Snapchat's explosively popular "female" filter (widely accepted to use a conditional variant of GAN), ends up consistently lightening the skin tones in women of color when trying to make face images appear more feminine. Our study is meant to serve as a cautionary tale for the lay practitioners who may unknowingly increase the bias in their training data by using GAN-based augmentation techniques with web data and to showcase the dangers of using biased datasets for facial applications.

Introduction

Breakthroughs in deep learning for image recognition have heralded significant progress in the field of computer vision, but one of the greatest limitations of the technology still remains: classifiers require massive amounts of training data to recognize meaningful patterns. As practitioners struggle to coax classifiers into generalizing to the underlying real-world distribution instead of overfitting to the sparse data, several data augmentation techniques have emerged as a useful form of regularization to increase training sample size and thereby, accuracy during test-time. The simplest of these augmentation techniques perform affine transformations on existing samples in the data, such as rotation, zooming, translation etc. (O'Gorman and Kasturi 1995; Bloice, Stocker, and Holzinger 2017). Ideally, the transformed samples should be representative of the same real-world distribution p_{data} as the original train and test sets. Thus, augmenting the training data with these samples should increase the likelihood of the classifier learning the real-world distribution p_{data}.

While in most cases the entirety of the Web 2.0 already produces the massive amounts of data necessary to solve deep learning problems, there are some instances where augmentation techniques are crucial. When relevant data is expensive to collect, as is the case for data behind a paywall, or continues to be scarce, as is the case for content which violates social media platforms' terms of service. Researchers rely on data augmentation for aggression detection (Aroyehun and Gelbukh 2018), for example, even as researchers are beginning the process of collecting sufficient datasets from social media platforms to benchmark models (Kumar et al. 2018). In recent times, the use of synthetic data has become popular among machine learning practitioners as an alternative technique for data augmentation (Teich 2019; Nisselson 2018). As the use of social media and web data for synthetic generation explodes in the AI community, it is imperative that lay practitioners be cognizant of the dangers and the limitations of using popular methods for synthetic data generation. Data collected from social media platforms or elsewhere on the web often carry the same social biases found in the real world. These latent biases can be picked up by a machine, even when these biases might not necessarily be obvious to us.

Generative Adversarial Networks (Goodfellow et al. 2014) can generate synthetic data by learning to mimic the style of the original, limited training set and create examples that seem novel (in contrast with affine transformations which may seem to practitioners to trivially duplicate the original train set). These examples give a (false) sense of sampling previously unseen data from the same underlying distribution as the original training data, which makes GAN seem like a promising candidate for data augmentation. We note that even this best-case scenario is still a territory for

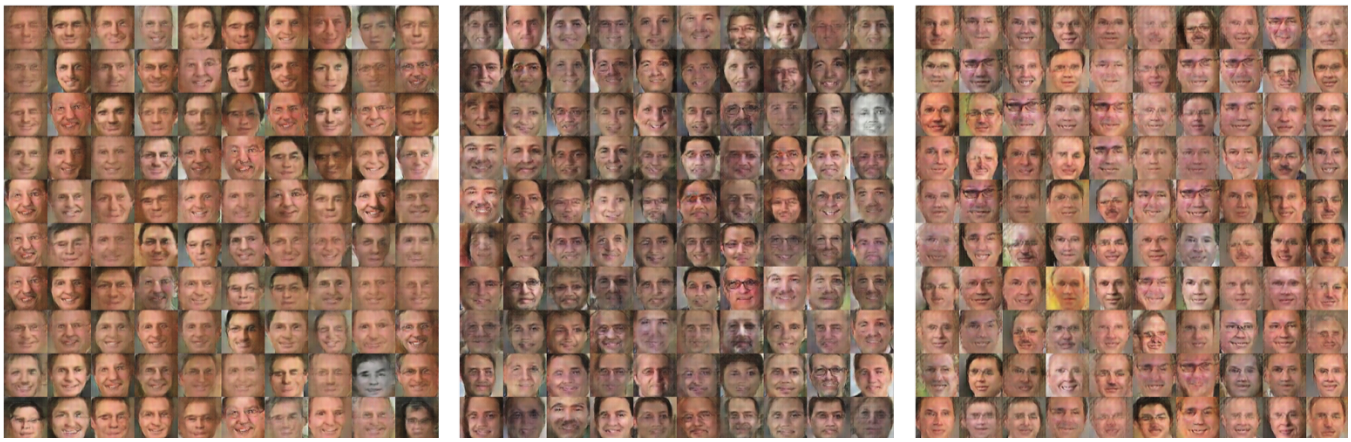

Figure 1: Synthetically generated engineering faculty images by the DCGAN after 50 epochs on three separate initializations. It is immediately apparent that the generator does not learn to create diverse skin colors. When human subjects were asked to label each image as having features related to gender or skin color, we noticed the presence of both gender and racial bias.

practitioners to tread lightly; GANs cannot be expected to learn better representations of the real world than any downstream classifier could, so augmented data will only propagate the existing biases of the real-world data.

However, GANs do not usually learn a good approximation of the real-world distribution (Arora and Zhang 2017). With human and machine classifiers, we show that the distribution they learn amplifies existing biases in the data, such as across gender and race. Data augmented from such a GAN will learn a distribution shifted from that of the real-world to pick up these exacerbated biases, disproportionately under-representing those already in the minority. If a training set fails to include minorities in a meaningful way, the AI systems that we rely on to make impartial, objective decisions will fail to consider minorities, as well.

We conduct an illustrative experiment to highlight how synthetic faces of engineering faculty are generated by a GAN (Fig. 1). We train the model on faculty headshots scraped from university directories of engineering schools across the country. It is unsurprising that the true distribution p_{data} is already non-uniform, with more white male professors than those of any other demographic, but we find that the GAN exacerbates the biases along the axes of gender and race. The generated images have even less representation for faces appearing to be non-white or female. For any downstream task, using a training data distribution which shifts away from and under-samples already underrepresented features has serious ethical implications, especially when biases exist along protected or embargoed attributes.

This failure mode of GANs impacts photo-editing applications already widely used today. We detail a preliminary study on a Snapchat filter meant to translate the input image of a face into one that is supposedly more feminine. This (presumably) GAN-based technology learns to adjust the colors of the pixels in the image and lighten skin tones of women of color to adhere to the female class. We find this performance is consistent across our examples for women of color using the technology, while it is not for white women. Although an exhaustive study on this effect with more data is necessary, our case study is the first to support the narrative of various users that Snapchat's selfie lenses lighten skin tones for women of color (Mulaudzi 2017; Jagota 2016; BetStyle 2016).

Architecture and Approach

The illustrative task of this work is to create a hypothetical augmented dataset of faces of engineering faculty and contrast it with the real-world dataset to highlight how this data augmentation technique perpetuates and even exacerbates racial and gender biases. In this section, we first review known limitations of GANs that are relevant to the problem of data augmentation. Second, we present and justify the use of a particular GAN architecture in our experiments with facial datasets. Finally, we detail the target distribution for the model to learn, and the process we used to collect our dataset for training and testing.

Model and Mode Collapse

GANs are known to solve the following minimax optimization problem between a generator network G and discriminator network D:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]$$

In the presence of infinite training data, computation time, and network capacity for the generator and the discriminator, this optimization will ensure that the distribution of the data generated by G converges to that of the training data (Goodfellow et al. 2014).

For our purposes, we assume that the discriminator D learns an approximation of the true distribution p_{data} given the finite set of training samples, from which images cor-

responding to faces of engineering researchers are generated. The generator G takes a random noise vector z sampled from some prior distribution p_z as input and generates a 64×64-pixel image $G(z)$. The hope is that D will not be able to distinguish between an image sampled from p_{data} and the image generated by G.

The first term in the optimization is proportional to the accuracy of D in classifying the actual data as real. The second term is proportional to the accuracy of D in classifying the synthetically generated images $G(z)$ as fake. Since the first term is independent of G, the discriminator aims to maximize the overall term while the generator aims only to minimize the second term. Over time, G learns the most important characteristics of p_{data} well enough to fool D to believe that the generated images are sampled from p_{data}.

While GANs do have theoretical guarantees for a tight approximation of their training distributions in the presence of infinite data, the scope of this paper is to examine the problems of GAN-based data augmentation where training data is limited, as one would not need to augment large datasets. It is important to analyze GANs in such realistic settings, and not only in theoretical ones. In this regard, we consider the task of generating facial images of engineering professors in the United States.

In (Arora and Zhang 2017), authors show that GAN-generated distributions p_{GAN} are not nearly as diverse as their training distributions p_{data}. The possible feature combinations in generated data are only representative of a small subset of what one can expect to see when sampling data from the real distribution. The authors explain in (Che et al. 2016), the generator G, which is a function of a low-dimensional vector of random noise z, collapses because the set of noise inputs that would correspond to some minority mode in the image space has a low or zero probability of being seen by G.

In other words, mode collapse is a direct consequence of the prior distribution of z. G rarely learns how to create images of these modes and is not penalized for missing them in the loss function, because the images it creates look real to D. While the distribution of random noise vectors is the primary factor of mode collapse, we predict that the skew of a non-uniform training set also determines the modes that the GANs will collapse to; our hypothesis is that for a highly unbalanced dataset, the modes G misses corresponds to a minority in the training set.

Although earlier works study the bias and generalization of GANs (Zhao et al. 2018), they do not discuss what happens to the generated distribution when the samples from p_{data} are non-uniform along certain dimensions. Furthermore, this work does not study the phenomenon of collapsing to modes, but rather to the opposite phenomenon of imagining unseen modes.

In this paper, we predict that for a dataset that is skewed along some axes (e.g. gender and skin color), G collapses to modes in the majority groups (e.g. masculine and white faces), thereby amplifying biases that exist in the original data. Intuitively, G is likely to exhibit this behavior because it may find it easier to fool D with features that commonly occur in the true distribution (eg. white skin color, males) than others that occur rarely.

We use the Deep Convolutional GAN (Radford, Metz, and Chintala 2015) (DCGAN) architecture. We choose to explore the question of how social biases are perpetuated by using this architecture for several reasons: First, it is a readily available off-the-shelf model that has been common among practitioners to use for data augmentation (Salehinejad et al. 2018; Frid-Adar et al. 2018; Guibas, Virdi, and Li 2017). Second, DCGAN uses the objective function mentioned above and thus, is susceptible to mode collapse. The authors of (Arora and Zhang 2017) show this is indeed the case, testing DCGAN on the CelebA datset of celebrity faces (Liu et al. 2015), CIFAR10 (Krizhevsky 2009), and the Bedroom dataset (Yu et al. 2015). This helps us understand the behavior of DCGAN on biased input data, which is the technical motivation for our study. Finally, it is the state-of-the-art model for face generation in the unconditional setting (DCGAN uses random noise as an input as opposed to starting with another image in the case of conditional GANs). Conditional GANs or their variants can generate highly photo-realistic images of faces (Karras, Laine, and Aila 2019), but they do not solve a problem of pure image synthesis. Conditional GANs such as pix2pix or CycleGAN (Isola et al. 2017; Zhu et al. 2017) do not generate images from random noise, but rather solves an image-to-image translation problem to adjust colors and textures in already-existing images to map it to another class. For example, one may transform a street scene to its segmented labels image, or translate an image of a horse to a zebra. StyleGAN, the current state-of-the-art for faces combines facets of image-to-image translation and style transfer (Gatys, Ecker, and Bethge 2015) by transforming input images into a final generated output. Our motivation is to explore the "imagined" distribution p_{GAN} and how the diversity decreases from that of p_{data} Since StyleGAN's output distributions are a combination of their input, we do not expect them to suffer the same lack of diversity as GANs generating images from random noise. We choose to study a GAN variant that solves the task of pure generation and uses the traditional objective function. For the remainder of this work, we will use the words "GAN" and "DCGAN" interchangeably unless stated otherwise.

Data Collection and Processing

To showcase the inherent biases that can exist in real-world data and how these are amplified by GAN-based data augmentation, we design a set of experiments around an already-biased dataset: faces of engineering professors collected from a selection of U.S. university webpages. This dataset reflects the biases among expert engineering academics in society; the web data samples a real-world distribution (unfortunately) biased toward masculine and white features. In the data collection process, we identified 47 of the universities listed in the most recent US News "Best En-

gineering Schools" ranking (U.S.N. 2019) that had public access to faculty directories with images. [1] We crawled the individual webpages of all engineering departments for each university to scrape the professor headshots. While parsing the headshots of professors, we omitted noisy input samples where logos or icons may obstruct the face of the person in some way. To do this, we tried to detect a face in the images we downloaded and automatically discarded those without one. The end result of this process was verified by a human. We exercised caution to ensure that the images belonged to faculty, and not non-teaching staff by (1) applying appropriate filters on the webpages before scraping and (2) handpicking images labeled as "staff" in the directory and removing them from our dataset. Using this process, our final dataset consists of 17,245 images of faculty faces across the country.

To ensure that the DCGAN learns features in the data relevant to the face, we perform regularizations to address two potential inconsistencies: (1) due to the fact that the majority of the faculty images are headshots taken in studio conditions, most of the backgrounds of the images in our training dataset were just of one solid color. We want to ensure that the generator will not pick up the trivial feature of background to fool the discriminator into classifying images as belonging to p_{data}, instead of learning to create realistic faces. (2) the photos vary: their dimensions differed from the 64×64-pixel input required by the discriminator and the faces do not have a consistent location across all images. It is imperative to establish a common ground across the data and diminish the trivial features and noise that could cause the DCGAN to underperform.

As a preprocessing step to address both problems, we crop all the headshots exactly to the face and scale them down to the desired 64×64 dimension. To do so, we use an unsupervised face detector based on *Histogram of Oriented Gradients* (Dalal and Triggs 2005). This method first renders the image in grayscale before dividing it into 16×16-pixel groups. It calculates average gradients for each group of pixels to match to a known representative face pattern. If it finds a match, it returns the coordinates of the group of pixels. We crop the image to the returned coordinates bounding the face and resizing it to 64×64 pixels.

Experiment and Results

In this section, we describe the studies conducted with human subjects and on a commercially available classifier to analyze the bias present in the original and the GAN-generated images along the axes of gender and skin color. We describe the experimental setup and analyze the results of the experiments: we show that the "imagined" faces of engineering faculty not only perpetuate the racial and gender bias present in the original data but also exacerbate it.

[1] We refrain from naming the specific universities in this write-up because we believe that the data is indicative of the presence of gender and racial biases among hired faculty in most universities in general as opposed to only in the ones we crawled.

Experimental Setup

Studies with Humans as Classifiers To analyze the potential gender and racial biases that the generator G of the DCGAN may have learned to fool D, we obtain human annotation on a set of randomly sampled images from both the original data and the generated data. To ensure we fairly assess the performance of the GAN, we obtain generated samples from three runs of the DCGAN initialized with different seeds. For each run, we use the facial images generated by DCGAN after 50 epochs. We then obtain human annotations on 50 images randomly sampled from each of the four sets of images (one real, denoted as x, and three synthetically generated, denoted as $G(z)$). [2]

We conducted four human study tasks, described as follows:

`T1a` Human subjects were asked to select the most appropriate option for an image x sampled from p_{data} with the following options: a) face mostly has masculine features, b) face mostly has feminine features, and c) neither of the above is true.

`T1b` Human subjects were asked to perform a task identical to `T1a`, but for a synthetically generated image $G(z)$.

`T2a` Human subjects were asked to select the most appropriate option for an image x sampled from the training data p_{data} from the list of following options: skin color is non-white, skin color is white, and can't tell.

`T2b` Human subjects were asked to perform a task identical to `T2a` but for a synthetically generated image $G(z)$.

To ensure that the data from those experiments was of high quality, we recruited 132 master Turkers on Amazon's MTurk[3] and paid each of them $1.20 for a seven-minute study. An MTurk worker with a master qualification is indicative of the fact that they have earned a high reputation by completing multiple tasks previously. We noticed that each Human Intelligence Task (HIT) took an average of five minutes to complete (while a couple of workers utilized the full seven minutes). Each worker was given a set of 52 images – 50 of which were from either the original dataset (x) or newly generated, i.e. $G(z)$ (but no mix-and-match), and two were high-quality images of celebrities – Scarlett Johansson and Idris Elba – for which the answers to all the tasks were trivial ('face mostly has feminine features' and 'face mostly has masculine features', respectively for `T1a` and `T1b` and 'skin color is white' and 'skin color is non-white', respectively for `T2a` and `T2b`). The last two images helped prune meaningless spam data generated by some of the human subjects who finished a HIT without actually paying attention to the question, or used a bot that gives a randomized answer to all the questions in the HIT. In our experiments, this filter helped us to identify and prune 18 such human subjects'

[2] We do not show the images randomly sampled from the original training set as they are images of real engineering professors who might not be comfortable disclosing their identity, especially in the context of our paper. However, we may share our dataset with researchers for academic purposes upon request.

[3] `https://www.mturk.com/`

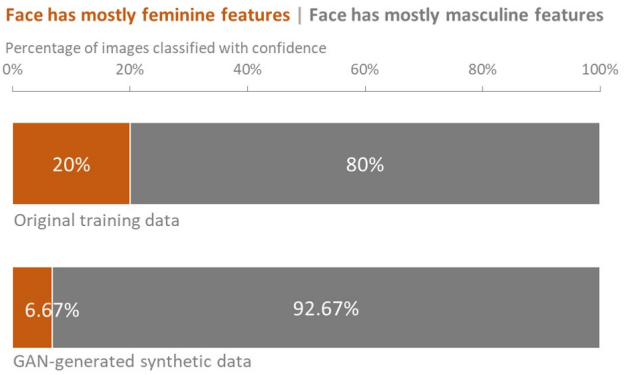

Figure 2: The percentage of faces classified as having feminine features by the majority of human subjects, decreased from 20% in the original dataset to 6.67% on average in the synthetically generated datasets.

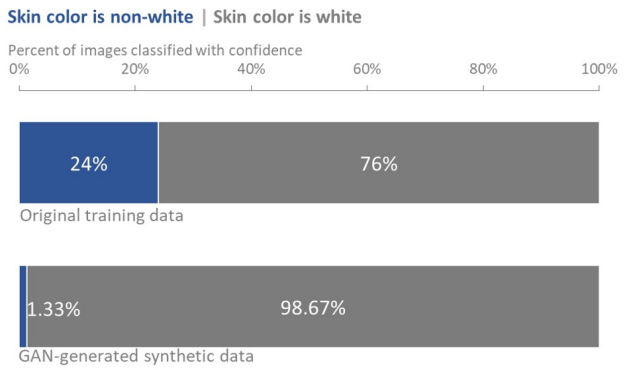

Figure 3: The percentage of faces classified as appearing non-white, by the majority of human subjects, decreased from 24% in the original dataset to 1.33% on average in the synthetically generated datasets.

answers. Nonetheless, we ensured that we obtained 30 valid data-points for each of the tasks, except for `T1a` for which we had 25 labels.

Studies with Commercially Available Classifiers To supplement the results obtained in the experiments done with human subjects, we also use commercial classifiers; this method replaces the human in the task of annotating the original and the generated images. On top of the three runs used to generate data for the human studies, we use two more runs of the DCGAN, initialized with different seeds, and a larger sample size of 1000 images (because annotation effort is cheaper in this automated process).

We use Microsoft Azure Cognitive Services' Face API[4] to detect the gender from a facial input image. We chose this classifier because it had the best overall accuracy among three commercially available classifiers for face data across the spectrum of gender and skin color (Buolamwini and Gebru 2018).

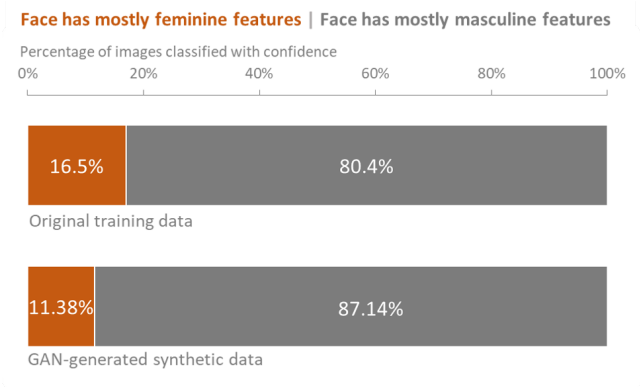

Figure 4: The percentage of faces classified as having feminine features by Microsoft Azure's Face API decreased from 16.5% in the original dataset to 11.38% on average in the synthetically generated datasets.

Analysis of Results

Bias in Original and Synthetically Generated Data We aim to assess whether the bias pertaining to the facial features associated with perceived gender and race in the initial dataset increases or stays the same in the synthetic dataset. For our analysis, we followed a majority-voting metric to categorize an image as having a feature. We plot the results for `T1a` and `T1b` in Fig. 2. Quite interestingly, we found the percentage of images that had mostly feminine features decreased from 20% in the original data to 6.67% in the generated data. A one-tailed two-proportion z-test on the original and averaged generated sample for the proportion of females assesses the null hypothesis that the proportion of feminine features in the synthetic distribution is equivalent to the proportion of those in the original distribution. This test yields a p-value of 0.0094, indicating statistical significance. We also observed a low p-value (6.4×10^{-5}) upon using Microsoft Azure's Face API for classifying input facial data to a gender. Images with feminine features decrease from 16.5% in the original dataset to 11.38% in the generated one (Fig. 4). Given the higher number of samples in this setting, this reduction in percentage turned out to be statistically significant (in support of the results we saw with human-annotated data). *These metrics show amplification of bias across the latent dimension of gender when a GAN is asked to 'imagine' an engineering faculty.*

We also noticed (as shown in Fig. 3), again using the concept of majority voting, for tasks `T2a` and `T2b`, the mode of images with non-white skin tones almost entirely vanished; the proportion of these images decreased from 24% in the original dataset to 1.33% in the synthetically generated dataset. The p-value obtained from the two-proportion

[4] `shorturl.at/bcQ01`

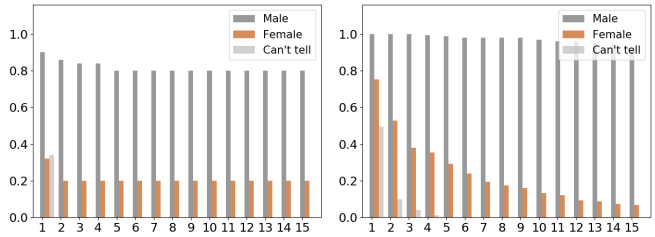

Figure 5: The number of images labeled as masculine, feminine, or neither, changes as the threshold number of votes required to categorize an image into a particular category increase from 1 to 15. Thresholding of the original and synthetic data are shown on the left and right, respectively.

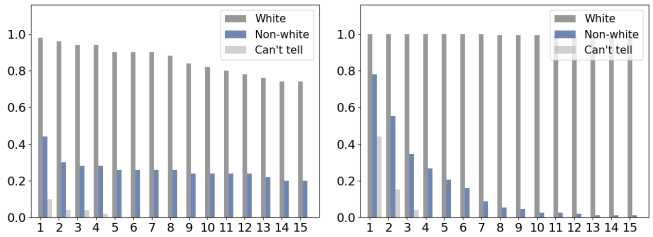

Figure 6: The number of images labeled as white, non-white, or can't tell, changes as the threshold number of votes required to categorize an image into a particular category increase from 1 to 15. Thresholding of the original and synthetic data are shown on the left and right, respectively.

z-test on the original and averaged generated sample proportion of non-white skin tones was 8.7×10^{-5}, showing strong statistical significance. *This result showcases that the GAN collapses along the latent dimension of race and learns to bias the synthetic faces toward lighter skin tones when asked to 'imagine' an engineering faculty*.

In the ideal, theoretical setting, the GAN should have generated data that preserves the bias seen on the real-world data. But in the realistic setting, we noticed that the situation is much worse: the synthetic data not only propagates but increases the biases against the minority populations in the true distribution, i.e. faces with feminine features and darker skin tones.

Confidence Metrics In Fig. 5 and 6, we plot the the fraction of total images (y-axis) for which at least n number of raters (x-axis) categorized as a particular option. In Fig. 5, the bar lengths for the various classes remain almost the same for the original data, indicative of the fact that the Turkers are more decisive and have higher confidence regarding which images have more masculine versus more feminine features. Given that the GAN-generated data is not photo-realistic, the Turkers become less decisive about the classification. Note that their confidence about which faces are more *feminine* decreases while their confidence about which faces are *masculine* almost stays the same. Thus, even for the faces classified as females based on our majority voting scheme, there is a large subset of Turkers that feel that these images have more masculine features.

Although the results for skin color follows a similar trend, we see even lower confidence among Turkers when classifying synthetically generated images (right) as non-white (Fig. 6). The percentage of images for which more than five (out of 30) Turkers believed the generated face is of a non-white person is below 20%.

Photo-Editing Filters: A Case Study on Snapchat

Though our experiment analyzed the lack of diversity only from DCGAN, conditional variants of GANs which perform image-to-image translation tasks are not exempt from mode collapse (Ma et al. 2018). We do not make claims for conditional GANs backed by careful studies as we do for DCGAN, but we still consider it crucial to discuss observations, implications, and further avenues for research when social media applications use any GAN which suffers from mode collapse, irrespective of its particular objective function. The effects of mode collapse on perpetuating and exacerbating social biases are not yet understood, and we wonder whether the effects have already led to unintended consequences in popular applications. We detail a preliminary, speculative study on Snapchat meant to serve as example.

In the context of photo-sharing platforms that exclusively provide filters to edit photos, Snapchat is the second-most popular network, after Instagram (Center 2019). Specifically, one of the most popular technologies core to the brand of Snap, Inc. is its computer-vision-assisted facial filtering, or "selfie lenses." A user can snap a picture with a dog nose and ears warped onto their face, which even moves with the user in real-time. Snap Inc. has continued to acquire computer-vision and artificial-intelligence startups to grow the scope and use of this technology. Last year, the multimedia messaging app released lenses that could be used to enhances a user's facial features to create a quintessential male or female in their gender selfie lenses. It seems the male filter makes the jawline more chiseled and adds facial hair while the female filter makes the chin pointed, slims the nose, and softens the facial features.

Snapchat has not officially released its dataset or the architecture specifications of its gendered lens algorithms, but it is believed in the machine learning community that image-to-image translation GANs to transform images of one class to another (such as CycleGAN) are a key component [5] (Magazine 2019; Jang 1970). Snapchat may have trained such a GAN to learn a mapping between two datasets of male and female faces to perform a "gender swap" task on input faces. The male and female selfie lenses are both available to all users. It remains an open question to explore how conditional variants of GANs react to sensitive social features, such as race and gender. Should the case be that the dataset

[5] shorturl.at/lMT89

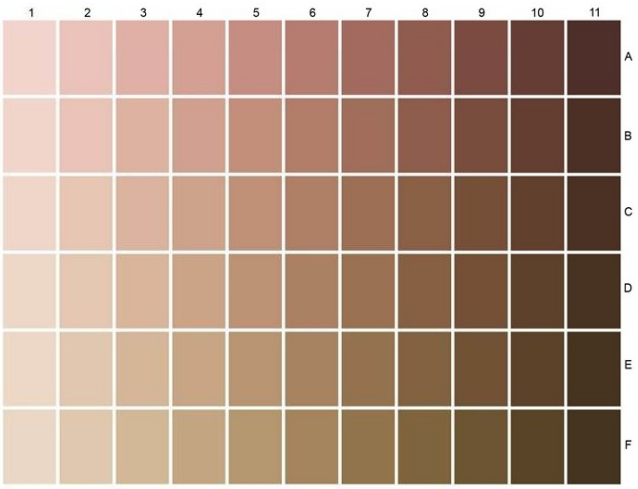

Figure 7: L'Oral skin color chart used for skin color identification. Lightness decreases from left to right and warmth decreases from top to bottom. [6]

is biased, any GAN will be susceptible to performing a one-size-fits-all translation, regardless of input.

The potential repercussions in the decision to release such filters to begin with (such as those regarding the sustaining of Euro-centric beauty standards for young users by way of changes to the harshness of facial features) aside, we speculated that the selfie lenses suffer from the same technical limitations our engineer-generating DCGAN did. And we do notice a dangerous relevant technical failure in the gender filters: the female filter appears to lighten skin tones for women of color. The essence of femininity is, apparently, not only soft features, but also a fair face, stemming from probable lack of diversity in the GAN-generated distribution.

A potential cause of the lightening of skin tones in women of color is that the mapping learned by an image-to-image translation GAN collapses to outputting light skin tones given any input of a female face. This indeed would be unsurprising given the performance of DCGAN on our dataset of engineering faculty, which learned to miss almost all non-white modes. While we have not performed a comprehensive study, the observation and claim open an intriguing research problem. Examples of the lightened complexions of women of color can be seen in Fig. 8, as contrasted with the changes when the filter is applied to the faces of white women in Fig. 9.

To qualitatively assess how the skin color changed between pairs of images, we manually cropped a section of the face under the eyes and above the tip of the nose, spanning both cheeks, found the average pixel value of this region, and mapped this RGB vector, using L2-norm distance, to the closest shades in the L'Oral color chart (Fig. 7) comprising 66 skin colors from around the world (LOral).[6] We did not consider the change in warmth, as shown within

[6] shorturl.at/mCST0

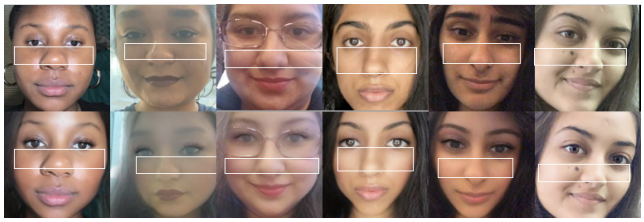

Figure 8: Faces of women of color before and after using Snapchat's female gender selfie lens, top and bottom, respectively. The sections used for the machine analysis are highlighted in white. Five faces are made one shade lighter by the filter. The second face's skin tone remains unchanged.

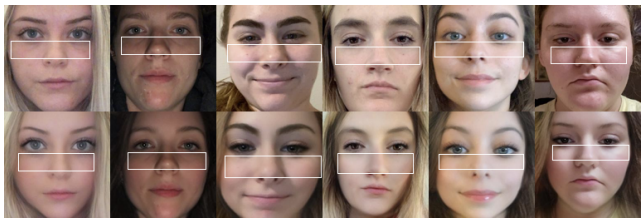

Figure 9: Faces of white women before and after using Snapchat's female gender selfie lens, top and bottom, respectively. The sections used for the machine analysis are highlighted in white. The first two faces are unchanged by the filter, the next two faces are made one shade darker, and the last two faces are made one shade lighter.

a column across rows; we only considered the change in the lightness. In other words, we discarded vertical shifts in the classes on the color chart. The selfie lens lightened non-white faces by one shade consistently for five faces and produced no effect for one face in our example. On the other hand, it acted randomly for white faces in our example, lightening two by one shade, darkening two by one shade, and not affecting two. An exhaustive study on this effect will require more data, but our case study offers initial support for the narrative of Snapchat's beautification selfie lenses lightening skin tones for people of color shared by various users across the web (Mulaudzi 2017; Jagota 2016; BetStyle 2016).

This phenomenon can be further studied if Snap, Inc. provides information on its GAN use or gives researchers insight into its data collection procedures. For example, if it uses its own users' facial images, it is likely its dataset is skewed toward white females, the single largest demographic of the application (this certainly would explain a mode collapse to light-skinned female faces). As opposed to being a final word on the matter, this case study seeks to open-up a research question about social bias perpetuation in social media (especially where GANs can be used) and engage thoughtful discussion.

73% of internet users aged 18-24 use Snapchat, and 51% of these Snapchat users are between 13 and 17 years old (Omnicore 2020). Teenagers and young adults are impres-

sionable, and their views of the world are shaped by what occupies most of their time. Today, social media and the web has the power to inject ideas into this generation which will influence them for the remainder of their lives. With over 250 million active monthly users under the age of 25 who spend an average of 40 minutes daily on the application, Snapchat holds the reins in controlling youth perceptions.

Colorism has afflicted communities of color for centuries; it has roots in European imperialism and Asian structures of hierarchy. Holding fair skin as a standard of beauty reflects traditional views that dark faces represent "savagery, irrationality, ugliness, and inferiority" (Hunter 2007). Colorism in the United States even continues to sustain multibillion-dollar cosmetic industries around the world. Without a complete understanding of colorism, racism, and classism, this dangerous societal standard negatively influences relationships of people of color and fosters self-hatred (Nittle 2019). Snapchat, one of the leading social media applications for users under 25 in the United States, ought not feed into this narrative for young women that light skin and femininity are synonymous. If anything, they have a social responsibility to release information on their data collection process and model architectures to allow for the community to conduct careful study on potential social bias.

Discussion

GANs have been shown to create less diverse data than the original data they are trained on. While it has been reasonably assumed in the community that this lack of diversity should be more drastic with non-uniform datasets (Antoniou, Storkey, and Edwards 2017), this phenomenon is not well-studied. It has also remained unclear what the implications of mode collapse are in scenarios where the training distribution p_{data} is biased toward certain feature values (eg. males) along a latent feature (eg. gender). To study this, we empirically show how a GAN trained on an image dataset of engineering faculty across the United States – which already is biased toward white and male professors – will pick up these inherent social biases and exacerbate them in the generated distribution p_{GAN}. In our setting, mode collapse occurs on a majority latent mode of the original data and causes a severe under-representation of feminine facial features and non-white skin tones in the generated dataset.

It is important that we do not teach our classifiers to capture this bleak lack of diversity that results from a technical failure mode of GANs.

Data augmentation with data generated using a GAN that utilizes the traditional training objective is faulty because it will suffer from the inevitable mode collapse. Though one would prefer the same distribution to underlie the training and augmented datasets, we show that this data augmentation technique causes a shift in the distribution toward biased extremes. Furthermore, GAN-based data augmentation aims to create large training data for today's data-intensive deep learning models, but as GANs are unable to capture the distribution of the real-world data, any downstream classification task relying on the biased data will be consequently biased.

The use of classification models on facial data is already prevalent in critical decision-making scenarios such as employment (Hymas 2019), healthcare (Bahrampour 2014), education (Kaur and Marco 2019), and criminal justice (Harwell 2019b). It is of clear ethical importance that we ensure our training sets are fair with respect to sensitive features. At the very least, they ought not to rig the system *against* already under-represented minorities. While it may seem to practitioners that GAN-based augmentation is more sophisticated than merely, say, rotating existing data, they are in reality using a technique that amplifies inherent biases in data across latent dimensions, potentially shifting the original distribution out of favor of minority groups. The implications of using a biased facial dataset for a classification task would be severe:

Classification Bias and Vulnerable Communities

1. *Minorities in Criminal Justice*: Classification systems for criminal justice, unfortunately, have already been identified as carrying bias against racial minorities. In 2016, ProPublica released a study that risk assessment software (Angwin and Larson 2016) used in state criminal justice systems is biased against black people. A classification system for recidivism risk given input with over 137 features – not including race – was found to disproportionately classify black defendants as medium or high risk, as well as rating their false positives as higher risk of re-offending than white defendants (This dataset was obtained from public records in Broward County available on the web – a detailed criminal dataset of over 18,000 individuals).

In the years since this study was released, facial-recognition systems have started to be used for policing (Harwell 2019b). Running surveillance footage through these systems can aid law enforcement officers to find and arrest suspects of crimes. The dangerous implications of under-representing minorities – or over-representing majorities – lies in the fact that African Americans and Hispanics are already disproportionately incarcerated, 5.1 and 1.4 times the rate of whites in state prisons, respectively (Nellis 2016), and criminal datasets used to train classifiers will reflect this. When these are used to train a GAN, the augmented data will further exacerbate the disproportionate representation of faces from these racial and ethnic minorities, perhaps almost completely ignoring white faces, similarly to how it performed for non-white faces with our dataset of engineering faculty. As it is difficult to remove race or ethnicity features for a facial classifier to consider, any criminal classification model using GAN-based architecture will be much more prone to marking black and Hispanic faces as criminals by default at a higher rate than white faces regardless of any crime being committed or not.

2. *Minorities in Employment*: While criminal justice systems were shown to exhibit racial bias before the use of facial data, employment systems were shown to exhibit gen-

der bias. In 2018, there was a public outcry over a hiring system designed (and scraped) by Amazon that showed bias against women (Dastin 2018). The goal of the classification system was to rapidly crawl the web for rsums in pursuit of likely good candidates. As it was trained on rsums of candidates over the past ten years, the classifier learned to exhibit the same gender biases existent in the male-dominated tech industry, despite never having been given gender as an input feature. It even learned to penalize rsums which showcased participation in women's extracurricular organizations or which indicated educational experience came from all-women institutions.

Since this debacle in AI-assisted employment technology, using facial data has became a critical part of hiring classification systems. Last year, HireVue facial expression technology was used for the first time to screen job candidates in the U.K., (Hymas 2019) and it has since been adopted by over 100 employers, analyzing over a million candidates (Harwell 2019a). HireVue announces that its technology is more favorable for diversity in hiring because it removes the possibility of human bias in hiring, such as if hiring managers only hire candidates who look and act like them. HireVue also claims to remove biased data if it causes an adverse impact on a protected class (hir 2019). But this is not always a feasible task as it is very expensive and not scalable; without extremely diverse and balanced datsets, models are likely to learn the patterns to bias against protected classes anyway. For example, in our experiment, skin color and gender were not given to the network (the model learns only from pixel intensities), but the GAN still exacerbated the biases it saw and missed creating non-white faces almost entirely. The employment-technology company has not released any information regarding the datasets it uses to train its algorithms or the specifications of the models it uses. With the likely assumption that it has been trained on previous candidates, it is susceptible to falling to the same fate of the failed Amazon rsum classifier; it will learn the social biases prevalent in the industries today and only favor those candidates who fit the status quo.

Deepfakes on Social Media

Recent advances in computer vision have made it easier to generate videos (as well as images) with very realistic, yet still synthetic, face imagery and make it more difficult for us to discern the real ones with our naked eye. This can raise implications on the trustworthiness of online information when used in fake news generation and identity theft in social media. Several works (Afchar et al. 2018; Güera and Delp 2018) aim to tackle these issues by training deepfake-detection models on face imagery datasets extracted from online video sources, but that fail to consider their potential gender and skin color biases. These include CelebDF (Yuezun Li and Lyu 2019) which scrapes millions of frames from 59 publicly available YouTube clips of celebrities and FaceForensics (Rössler et al. 2018) with a half-million frames extracted from 1,004 YouTube videos. One face imagery dataset which does not take from the web, DeepFake-TIMIT (Korshunov and Marcel 2018), amasses a dataset of 320 face-swapped videos with 16 similar-looking pairs of people. None of these works check for fairness over the gender or skin-color attributes; hence, they can pose imbalances toward the most (or lesser) prominent categories when used to train the aforementioned detection models. A better starting point would be to use a dataset with provisions already in place to ensure fairness with respect to skin-color and gender, such as created by Facebook in its Deepfake Detection Challenge (DFDC) Preview Dataset (Dolhansky et al. 2019). Participants in this competition may leverage GANs to build systems that match generated images with the potentially fake social-media posts, but this would even further increase the inherent biases in their data, leading to potential dangerous false-positives at an inconvenience to social-media users.

Further Implications

Beyond implications about social issues, this work should serve as a general caution against using GAN-based data augmentation techniques for any downstream task such as in medical imaging or detection of chip defects. There seems to exist a false sense of security that GANs can generate *novel* data samples which pick the expected semantic features relating to the defect and place them in previously unseen settings. The augmented data might be under-representing some crucial feature of the real-world data, since we show that mode collapse is a tangible problem even in latent dimensions. Our AI systems do not take in gender and race as input features, yet they learn to skew generated samples along those axes anyway. Practitioners must understand that while their data might seem balanced to the human eye, there likely exists a skew across some hidden feature which will be picked up by a GAN.

Conclusion

The bias of yesterday on gender or race influences the decisions of today and tomorrow. The use of such data for augmentation in present day classification systems will make the future challenge of correcting biases arduous as it will have to offset more data that amplifies the bias. Lastly, we note that while we focused on perpetuation of obvious and troubling social biases, the caution should be extrapolated to other fields; GAN-based data-augmentation can perpetuate any type of extraneous bias. This should give pause to the widespread use of GAN-based data augmentation in medical and anomaly-detection domains, perhaps in favor of reliable and well-understood data augmentation techniques.

Acknowledgments

Kambhampati's research is supported in part by ONR grants N00014-16-1-2892, N00014-18-1-2442, and N00014-18-1-2840, AFOSR grant FA9550-18-1-0067, NASA grant NNX17AD06G and a JP Morgan AI Faculty Research grant.

References

[Afchar et al. 2018] Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. *CoRR* abs/1809.00888.

[Angwin and Larson 2016] Angwin, J., and Larson, J. 2016. Machine bias. https://bit.ly/37uNrV2.

[Antoniou, Storkey, and Edwards 2017] Antoniou, A.; Storkey, A.; and Edwards, H. 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

[Arora and Zhang 2017] Arora, S., and Zhang, Y. 2017. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*.

[Aroyehun and Gelbukh 2018] Aroyehun, S. T., and Gelbukh, A. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 90–97.

[Bahrampour 2014] Bahrampour, T. 2014. Can your face reveal how long youll live? new technology may provide the answer. https://wapo.st/2GrQKjL.

[BetStyle 2016] BetStyle. 2016. Hold Up: Does Snapchat Have a Problem With Brown Skin? https://bet.us/3aJ42qb.

[Bloice, Stocker, and Holzinger 2017] Bloice, M. D.; Stocker, C.; and Holzinger, A. 2017. Augmentor: an image augmentation library for machine learning. *arXiv preprint arXiv:1708.04680*.

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

[Center 2019] Center, P. R. 2019. Social media fact sheet. https://pewrsr.ch/2U9gs4X.

[Che et al. 2016] Che, T.; Li, Y.; Jacob, A. P.; Bengio, Y.; and Li, W. 2016. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*.

[Dalal and Triggs 2005] Dalal, N., and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *In CVPR*, 886–893.

[Dastin 2018] Dastin, J. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. https://reut.rs/2GlVckf.

[Dolhansky et al. 2019] Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

[Frid-Adar et al. 2018] Frid-Adar, M.; Klang, E.; Amitai, M.; Goldberger, J.; and Greenspan, H. 2018. Synthetic data augmentation using gan for improved liver lesion classification. In *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on*, 289–293. IEEE.

[Gatys, Ecker, and Bethge 2015] Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

[Goodfellow et al. 2014] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

[Güera and Delp 2018] Güera, D., and Delp, E. J. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–6. IEEE.

[Guibas, Virdi, and Li 2017] Guibas, J. T.; Virdi, T. S.; and Li, P. S. 2017. Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*.

[Harwell 2019a] Harwell, D. 2019a. A face-scanning algorithm increasingly decides whether you deserve the job. https://wapo.st/30VvdcD.

[Harwell 2019b] Harwell, D. 2019b. Oregon became a testing ground for amazons facial-recognition policing. but what if rekognition gets it wrong? https://wapo.st/2TXiXHl.

[hir 2019] 2019. Ai-driven pre-employment assessments for recruiters and ta. https://www.hirevue.com/blog/creating-ai-driven-pre-employment-assessments.

[Hunter 2007] Hunter, M. 2007. The persistent problem of colorism: Skin tone, status, and inequality. *Sociology Compass* 1(1):237–254.

[Hymas 2019] Hymas, C. 2019. Ai used for first time in job interviews in uk to find best applicants. https://bit.ly/2vmBwu9.

[Isola et al. 2017] Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

[Jagota 2016] Jagota, V. 2016. Why Do All the Snapchat Filters Try to Make You Look White? https://www.complex.com/life/2016/06/implicit-racial-bias-tech.

[Jang 1970] Jang, E. 1970. Fun with snapchat's gender swapping filter. https://blog.evjang.com/2019/05/fun-with-snapchats-gender-swapping.html.

[Karras, Laine, and Aila 2019] Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

[Kaur and Marco 2019] Kaur, H., and Marco, T. 2019. A new york school district is bringing in facial recognition software. rights groups say it could spell trouble for students. https://cnn.it/30WLaiX.

[Korshunov and Marcel 2018] Korshunov, P., and Marcel, S. 2018. Deepfakes: a new threat to face recognition? assessment and detection. *CoRR* abs/1812.08685.

[Krizhevsky 2009] Krizhevsky, A. 2009. Learning multiple layers of features from tiny images. Technical report.

[Kumar et al. 2018] Kumar, R.; Ojha, A. K.; Malmasi, S.; and Zampieri, M. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop*

on Trolling, Aggression and Cyberbullying (TRAC-2018), 1–11. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

[Liu et al. 2015] Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[LOral] LOral. Expert in skin and hair types around the world - L'Oral Group. https://bit.ly/2tQhOGI.

[Ma et al. 2018] Ma, S.; Fu, J.; Wen Chen, C.; and Mei, T. 2018. Da-gan: Instance-level image translation by deep attention generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5657–5666.

[Magazine 2019] Magazine, P. 2019. The dark implications of facial swap filter technology. https://bit.ly/2vrmjbv.

[Mulaudzi 2017] Mulaudzi, S. 2017. Let's Be Honest: Snapchat Filters Are A Little Racist. https://bit.ly/2NXCJhR.

[Nellis 2016] Nellis, A. 2016. The color of justice: Racial and ethnic disparity in state prisons.

[Nisselson 2018] Nisselson, E. 2018. Deep learning with synthetic data will democratize the tech industry. https://tcrn.ch/2RRUAYP.

[Nittle 2019] Nittle, N. K. 2019. The harmful effects of colorism. https://www.thoughtco.com/the-effects-of-colorism-2834962.

[O'Gorman and Kasturi 1995] O'Gorman, L., and Kasturi, R. 1995. *Document image analysis*, volume 39. IEEE Computer Society Press Los Alamitos.

[Omnicore 2020] Omnicore. 2020. Snapchat statistics. https://www.omnicoreagency.com/snapchat-statistics/.

[Radford, Metz, and Chintala 2015] Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[Rössler et al. 2018] Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR* abs/1803.09179.

[Salehinejad et al. 2018] Salehinejad, H.; Valaee, S.; Dowdell, T.; Colak, E.; and Barfett, J. 2018. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 990–994. IEEE.

[Teich 2019] Teich, D. A. 2019. Synthetic data is a tool for improving training and accuracy of deep learning systems. https://bit.ly/36moolG.

[U.S.N. 2019] U.S.N. 2019. Best Engineering Schools. https://www.usnews.com/best-graduate-schools/top-engineering-schools/eng-rankings.

[Yu et al. 2015] Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR* abs/1506.03365.

[Yuezun Li and Lyu 2019] Yuezun Li, Xin Yang, P. S. H. Q., and Lyu, S. 2019. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*.

[Zhao et al. 2018] Zhao, S.; Ren, H.; Yuan, A.; Song, J.; Goodman, N.; and Ermon, S. 2018. Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems*, 10792–10801.

[Zhu et al. 2017] Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.