Skewed EleGANce: Implications of CycleGANs Aggravating Biases on Facial Data

Fiona Peng Dr. Suleyman Uludag

9 September 2022

Abstract

Generative Adversarial Networks (GANs) have become a widely-used data augmentation technique for data-demanding deep learning models. In this paper, we demonstrate that CycleGAN, a popular image-to-image translation GAN, not only perpetuates, but aggravates biases along the axes of skin tone and gender when given a skewed training dataset of face images. While CycleGANs can synthetically generate unique, unseen data samples, they are not as elegant as they appear due to mode collapse, which is when the generators collapse and are only able to produce a small variety of samples. Their use can lead to adverse effects when applied to real-world training datasets that are biased along latent dimensions. We empirically demonstrate that the GAN-generated dataset displayed a significant under-representation of non-white skin tones and feminine facial features when stylizing representative faces to look like the faces of members of technology companies. To demonstrate real-world implications of mode collapse in CycleGANs, we perform a case study on Snapchat's CycleGAN-based "My Twin" filter, which disproportionately lightens the skin tone of women of color. This work acts to caution the use of GAN-based data augmentation methods in downstream tasks, which can lead to adverse consequences by further exacerbating biases against minority groups. It is critical that from an ethical standpoint, GAN practitioners for data augmentation ensure that both training datasets and trained models are representative and diverse regarding sensitive features. We also recommend some future work.

Keywords: Generative Adversarial Networks (GAN), CycleGAN, Data augmentation, Mode collapse, Facial data, Image-to-Image Translation, Aggravation of bias

Contents

1 Introduction

Data availability has become an increasing challenge with the rise in deep learning in our modern society driving significant advances in numerous separate fields. Deep learning models typically require a substantial amount of data for training in order to obtain high predictive performance, which can be a problem when there is limited data due to the high cost for data acquisition, privacy and confidentiality concerns, paywalls, or challenges in labeling data. Because of this limitation in data availability for deep learning, practitioners and researchers have turned to reliable data augmentation techniques, one of the most promising being the use of Generative Adversarial Networks (GANs) [1], which are models that create new data instances resembling the training data. GANs play a crucial role in data augmentation in several fields, including generating synthetic data for medical imagery [2] and facial photos [3]. GANs achieve this through a generative neural network model that consists of two network components: a generator focused on generating synthetic images, and a discriminator focused on discrimination among the images. Essentially, it is approximating the original distribution with a given limited dataset to create original and unique data. Along with GANs, CycleGANs [4], which are conditional GANs (cGANs) [5] involved in image-to-image translation, have also been popularly used to augment datasets. In this paper, we focus primarily on data generation using this specific variant of GAN, CycleGAN. Using CycleGANs can generate data that appears to be novel from the same distribution of training data.

In addition to data augmentation, GANs have also shown significant potential in other tasks, such as the generation of other unstructured data, including natural images, image super-resolution, video prediction, and 3D object generation [6] [7]. As a result, the use of GANs has increased significantly in recent years as research and implementation of deep learning models has also increased. GANs have significant potential, but it is important to note that they are vulnerable to mode collapse [6][8], which is when the generators collapse and only generate a small set of outputs, called modes, due to these outputs being able to easily fool the discriminator. As a result, GANs have limitations regarding bias propagation and aggravation, which is the focus of this paper.

1.1 Face Data Augmentation With GAN

This paper focuses on facial data augmentation: generating synthetic images of face data. Human faces play an essential role in personal identification, interaction, and emotional expression . Similar to other deep learning models, the quality and amount of training data is critical [3]. Face data augmentation is a technology that aims to increase the size of datasets for model training and testing through transforming real data samples. Face data augmentation is significant in maximizing the performance of deep neural networks because: (1) It is significantly less expensive to generate synthetic data that includes annotations than manually labeling and collecting real data. (2) Synthetic data protects the privacy of individuals and allows for a balanced dataset. (3) Synthetic augmented data by GANs can be extremely accurate. (4) Faces that include specific attributes and features can be generated using augmentation techniques [9].

However, it is important that practitioners of these GANs for generating synthetic data images note their limitations. GAN-generated data for the goal of data augmentation would only aggravate the existing irregularities and bias that are present in real-world data. Regarding facial image data, this bias could include irregularities in gender, race, ethnicity, and prominence of certain features in the training dataset. Taking into account the limitations of GANs [11], it is crucial that researchers work with caution. Synthetically-generated data from GANs would learn a shifted distribution– one that is not representative of real-world data– which only further intensifies the biases that are already prevalent in society. In other words, the use of GANs would disproportionately under-represent, both in quality and number, populations that are already in the minority by amplifying the skew of the dataset. Machine learning models used on facial data are currently widespread in important decision-making applications, including healthcare [12], employment [13], criminal justice [14], education [15], and security developments such as deep-fake detection [16]. As a result, severe ethical implications are brought to attention, especially when bias on synthetic GAN-generated data exists on protected attributes, which may lead to

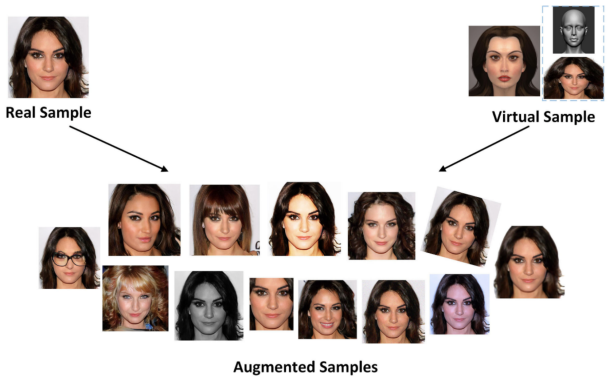

Figure 1: Illustrative diagram of face data augmentation. In this example, the augmented samples are generated by TL-GAN [10].

unintentional targeting of certain minority groups. This leads to harmful consequences for these groups when GAN-based generation using a skewed dataset is applied to real-world scenarios, as further discussed in Section 6 of this paper.

1.2 Contribution of This Work

To our knowledge, we are the first to empirically analyze how CycleGAN intensifies biases along the lines of skin color and gender when provided a skewed, unrepresentative distribution of face-shots of the executive and board team members of the top-100 technology companies [17]. Our hypothesis is that for a dataset that is already biased and skewed along latent (inferred) axes, such as skin tone and gender, the generator G of the CycleGAN model will collapse to modes in majority groups, which in this case are white and masculine faces. We predict the result of this mode collapse will be the amplification of existing bias in the original training dataset (Figure2).

Figure 2: (a) Samples of original images from the representative Chicago Face Database [18], and (b) Samples of CycleGAN-generated images from representative Chicago Face Database images to faces of members on executive and board team of top-100 technology companies [17].

Next, we perform a case study on a real-world application of CycleGAN, namely, for Snapchat's gender-swap and gender-blending filter. According to several sources [19][20][21], this filter is based on a CycleGAN framework, so we study potential biases when applying this filter to fe-

male versus male faces, and faces of varying skin tone. We show that this CycleGAN-based filter reacts to the sensitive features that are discussed in this paper.

The rest of the paper is organized as follows: First, we present a short background on GANs, CycleGANs, and potential applications. Then, in Section 3, we review related works and contrast them to ours. In Section 4, we present the datasets we used and how we processed that data. We also discuss our training environment. In Section 5, we share our experiment with CycleGAN on biased data and evaluate how data augmentation with the model propagated irregularities. We also include a case-study on the use of CycleGANs in Snapchat gender-swap filters. In Section 6, we discuss our findings and connect it to real-world applications. Finally, in the last section, we conclude our findings and highlight future work.

2 Background

In this section, we provide a background on GANs, including its framework and applications. We also discuss CycleGAN, an image-to-image translation variant of GAN, and contrast it with previous image-to-image translation models that require paired data. We also provide a summary of popular real-world applications of CycleGAN, and how they are used in facial data augmentation. Finally, we discuss a major limitation of GANs: mode collapse.

2.1 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) were first proposed by Goodfellow *et al.* [1]. The goal of GANs is to generate synthetic images through training from input data. Essentially, after the model is trained, the network will learn to be able to generate a new output image through mapping from a random noise vector (e.g., Gaussian or uniform), which is a feature vector that is unique for every image. GANs consist of two main components: a generator network G and a discriminator network D. Both G and D typically adopt the structure of the widespread deep convolutional neural networks that are popularly used today [7].

The generator network, G, is able to generate synthetic images, while D is a binary classifier that is able to predict if the image input is a "real image" that was a target image originally present in the dataset, or a "fake image" that was generated by G. The GAN framework is illustrated in Figure3. As D learns to discriminate between images that are generated by G and images that come from a real-world dataset distribution, G tries to generate fake image outputs that are able to fool D into incorrectly guessing that the generated image is a real image.

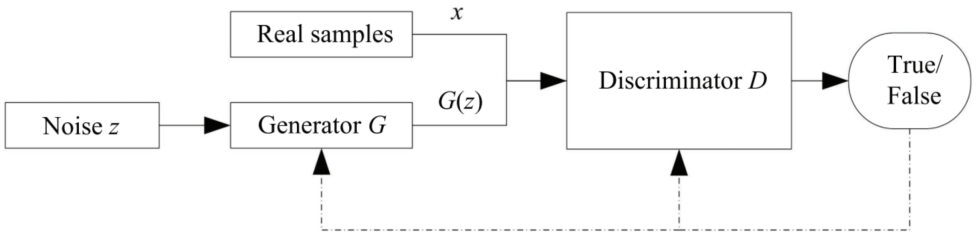

Figure 3: The GAN framework structure, which simultaneously trains the generator and discriminator networks. [1][22]

G and D eventually "compete" in a minimax game against each other to get better than the other, leading to both consistently improving to "beat" the other. G strives to generate fake images that are so similar to the real images that they're able to fool D. At the same time, D strives to constantly alter its weights to be able to better classify the input images from G; in other words, it is trying to be able to "catch" the fake images from G and label them as fake to avoid being fooled. When unlimited training data, network time, and computation time are present, this process continues until the generated images from G are so similar to the original

target image that they're practically indistinguishable by D (they have converged), meaning that the generated adversarial images are very realistic [23].

GANs have achieved very impressive visual results in the fields of image generation, representation learning, and image editing [22]. They have become a popular research topic as artificial intelligence is booming. In regard to data augmentation through synthetic image generation, earlier research surrounding computer vision was directed toward generating more data through affine transformation to established sample data [24]. Alternatively, GANs have become a popular mechanism for generating synthetic data because they use a different approach that is able to generate data that looks to be novel [25]. GANs have also been widely applied to generate photorealistic objects like faces. Recent methods have also adopted the framework of GANs for conditional image generation applications using conditional GANs (cGANs) [5], including image inpainting, text2image, and future prediction.

2.2 CycleGAN

CycleGAN [4] is a technique that uses an extension of the GAN architecture to automatically train a deep convolutional neural network [7] to accomplish image-to-image translation tasks. Image-to-image translation is a cGAN that involves generating fake forms of an input image with specific transformations, such as translating black and white images to color, or transforming summer photos to winter photos. According to Zhu *et al.*, "Image-to-image translation is a class of vision and graphics problems where the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs" [4]. Using a pair of unrelated datasets, CycleGAN is able to learn mapping between input and output images, such as after a transformation is applied. These models utilize unsupervised training by using training images from the source and target images that do not need to be paired, or in other words "related". This is a very powerful technique and is able to achieve impressive results for a variety of applications, such as translating photos of horses to zebras, or even regular photographs to Monet paintings.

Traditionally, it was a requirement for image-to-image translation models to use a training dataset that includes paired data, as is the case in the popular image-to-image translation model, pix2pix [26]. Paired data models require the training dataset to include two collections, containing the same images, but one collection having transformed images of the other. This is a limitation because it can be very challenging and expensive to acquire and prepare these photos. CycleGAN, the model evaluated in this paper, is a method for unpaired image-to-image translation. CycleGAN works on two collections of unrelated images by extracting certain characteristics from each to be used in the image-to-image translation process. For example, it could take unpaired data such as a large collection of photographs of summer scenery and a large collection of photographs of winter scenery with unrelated landscapes, and the model will be able to translate photos between the two groups, whether that is summer landscape to winter landscape photos or winter landscape to summer landscape photos.

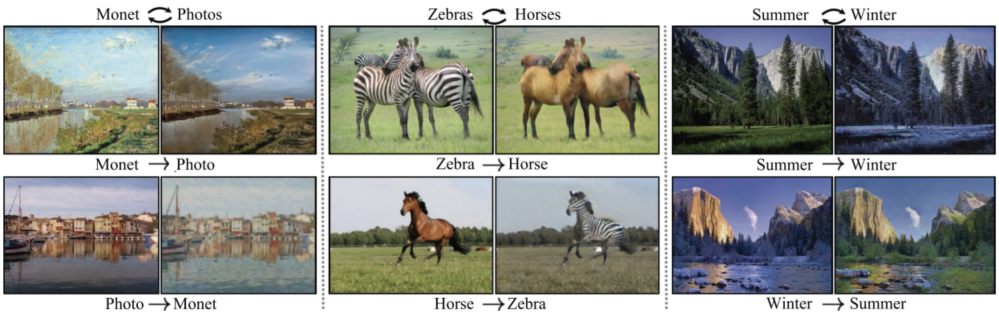

Figure 4: Example CycleGAN image-to-image translation results for unpaired image-to-image translation, used on various applications. [4]

CycleGANs follow a similar architecture to traditional GANs and also consist of a generator network G and a discriminator network D. However, unlike GAN or pix2pix, CycleGAN extends

the GAN architecture and trains two generator networks (G_1 and G_2) and two discriminator networks (D_1 and D_2) simultaneously, as shown in Figure5. One generator (G_1) inputs images from the first domain and translates them to produce output images for the second domain, while the other generator (G_2) inputs images from the second domain and translates them to produce output images for the first domain. Likewise, the two discriminator networks then determine how realistic the synthetic images look and update G_1 and G_2 accordingly.

2.2.1 Cycle Consistency

Although the extension of using two generator networks and two discriminator networks may be able to generate accurate images in both domains, by itself it is not able to generate translations of the input images [27]. The adversarial losses can't guarantee that the function can correctly map and translate an input image to the desired output image. To do this, CycleGAN implements an extension to the original architecture, called cycle consistency [28]. Cycle consistency is the concept that if we translate a sentence from Language A to Language B and then translate it back from Language B to Language A, the output sentence should match the original sentence [29]. In terms of GANs, cycle consistency is the idea that the generated image from G_1 could be taken in as input for G_2. Thus, a generated output image from G_2 should match the original image that G_1 took as input. Additionally, the reverse should also comply: the output image generated by G_2 can be taken in as input for G_1, and the results should also match the original image that G_1 took as input.

CycleGAN allows for cycle consistency through incorporating an additional loss to measure the offset between the GAN-generated output of G_2 and the original image given as input to G_1, and the reverse, e.g. using the summed absolute difference or L1 norm in pixel values [29]. This is designed to take into account and encourage the generated target images that are image-translations of the input image. Essentially, this regularizes the two generator networks and eventually guides the image generation process toward image translation.

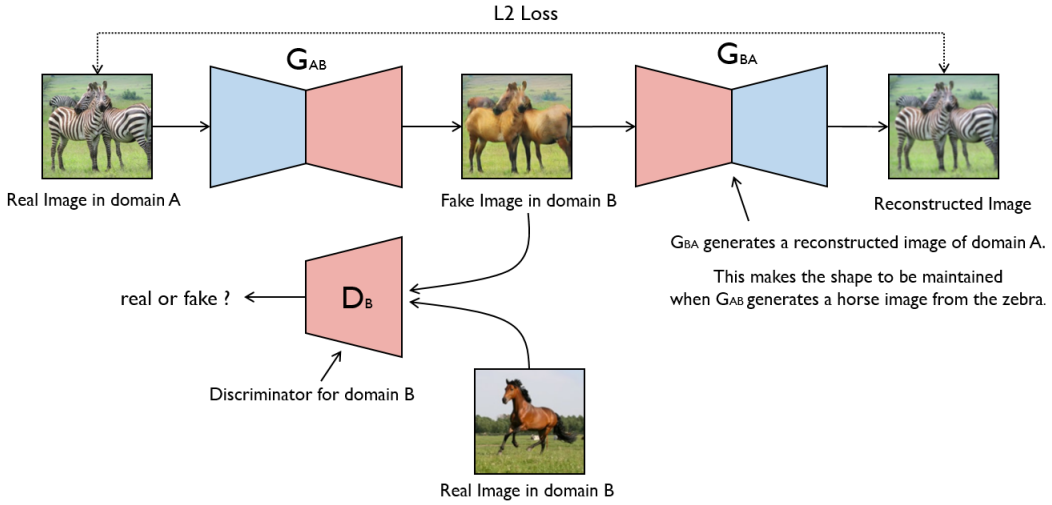

Figure 5: CycleGAN framework; (1a) G_1 takes photos of zebras (collection 1) as input, and outputs synthetic photos of horses (collection 2). (1b) D_1 takes original dataset photos of horses (collection 2) and also the output of G_1 as input, and outputs the probability that the image is from collection 2. (2a) G_2 takes photos of horses (collection 2) as input and outputs synthetic photos of zebras (collection 1). (2b) D_2 takes original dataset photos of zebras (collection 1) and also the output of G_2 as input, and outputs the probability that the image is from collection 1. [30]

2.2.2 Applications of CycleGAN

CycleGAN is one of the most widely-used GAN variants among machine learning researchers and practitioners, according to the number of stars on their most-used GitHub repository [31]. CycleGANs have a variety of successful applications. For example, they can be used to convert images of horses to zebras, and vice versa, aerial photos to Google maps photos, and summer landscapes to winter landscapes (Figure4). CycleGANs are also widely used for data augmentation to address data scarcity. A critical real-world application of cGANs, including CycleGANs, is in the domain of medical imagery, where one could use GANs to augment rare data of unhealthy patients through injecting or removing different regions of the image [2][32][33].

CycleGAN can also be used on facial image data for augmentation, or for swapping one's gender or editing images to have certain attributes [9][34]. Generative models, such as the transparent latent-space GAN (TL-GAN) [10], are a powerful tool for generating new data by learning the the training dataset's distribution. Figure1 shows a schematic representation of face data augmentation with TL-GAN. Since the general GAN concept was proposed, profound research has been done to improve its capability and stability. Phung *et al.* [35] presented a face data augmentation framework using CycleGAN. The authors also compared performance of CycleGAN to that of the classic GAN. From the experiment, they were able to conclude that though the classic GAN learned mapping from low-dimensional manifold to high-dimensional data domains (images), CycleGAN was able to learn the mapping between two high dimensional data spaces. Thus, CycleGAN is more efficient than the original GAN in completing and complementing an imbalance dataset [35].

2.3 Mode Collapse

Although data augmentation through GANs appears to be the perfect solution to the ever-pressing challenge of data scarcity for deep learning models, in reality, GAN-generated data is not nearly as diverse as their training data counterparts [36][11]. A known challenge of GANs, including CycleGANs, is that they are difficult to train because of their sensitivity to training instability and hyper-parameters [6]. On the other hand, another very common pattern of failure during training is when the generator can only produce one or a small set of outputs, or modes. This is called mode collapse [23]. When this happens, the generators are still generating meaningful samples, but these samples only include a couple modes, which typically are small regions in the data distribution that have a higher-than-usual probability because they are able to easily fool the discriminator. This phenomenon is caused by the "missing modes problem" [6], which is considered a severe problem during GAN training. Specifically, many modes in the original distribution are not represented in the generated samples, and the support size of these synthetic images is restricted by the capacity of the discriminator D. This causes a significantly lower entropy distribution that has less variety than the original data generating distribution. Due to the substantial consequences of this issue, many recent research papers have addressed this challenge and proposed strategies and novel architectures that are able to stabilize the training of these GAN models, leading to more sample diversity, as discussed further in the upcoming Related Work section.

3 Related Work

This section explores several related works that study generation and bias in GANs to unseen and existing modes, as well as research surrounding mode collapse. We also discuss several related works that study implications of bias propagation presented in GANs, and contrast these works to our own.

There are related works that study GANs' generation and bias to unseen modes. Zhao *et al.* [8] researched the inductive bias of deep generative models, including GAN, and proposed a framework to demonstrate both bias and generalization in state-of-the-art GANs. Through their experiments, they were able to discover patterns that are consistent across models, hyper-parameter choices, and datasets, and even indicate surprising similarities with previous experiments within the field of cognitive psychology. However, they don't address GANs' problem

of collapsing to existing modes. While there is widespread research being performed on mode collapse ([23][37][11]), and there are a couple variant GANs with altered frameworks that have been implemented and developed to mitigate the effects of mode collapse [38][39][40], there has not been much distinction made between non-uniform and uniform training datasets.

Along those lines, Mishra *et al.* [39] was able to show empirically that there is indeed a contrast between the original training dataset and the GAN-generated dataset. They also showed that this divergence worsens as the distribution of the training dataset worsens and becomes more skewed. They were able to accomplish this through training a latent space inversion network with GAN using a divergence loss. Mishra *et al.* then validated their approach method on various tasks such as conditional generation and mode separation using real-world datasets. In the end, they were able to demonstrate its efficiency over other state-of-the-art models. However, Mishra *et al.*'s research was focused on using four scalar metrics, which unfortunately does not provide adequate information on the question of how the distributions actually differ.

Several other works also discuss the implications of the bias propagation presented in GANs. Jain *et al.* [41] researched how GANs are able to propagate existing bias when they use an already-skewed dataset of head-shot photos of engineering schools' faculty members. Through their experiments, they were able to demonstrate that traditional GANs– (1) DCGAN (the most common GAN used by researchers and practitioners) and (2) ProGAN (a state-of-the-art GAN that is known to not only address the quality-variance trade-off, but also the mode-collapse problem)– further distort and decrease diversity in the distribution of a dataset that contains headshots of engineering faculty at various schools.

Like Jain *et al.*, Kenfack *et al.* [36] also researched representation bias and unfairness in GANs, emphasizing that fairness has become an essential problem in not only GANs, but various other machine learning domains, including natural language processing and classification. They defined a novel fairness notion (representational fairness) for GANs that shows the degree of similarity between generated samples sharing the same protected attributes, such as race, gender, etc. They showed that this fairness notion is violated even in the ideal case when the training dataset consisted of an equally represented group, showing that the generator favored generating one group of samples over others, demonstrating the phenomenon of mode collapse. Specifically, in Kenfack *et al.*'s work, they propose a method to overcome this problem by controlling the groups' gradient norm. They accomplished this through performing gradient clipping in the discriminator during training, and their findings show that this approach generated data that is more fair with regard to representational fitness. This addressed the knowledge gap in how the distributions actually differ that Mishra *et al.* did not thoroughly address.

These works contrast to our work in the following ways:

(1) We are the first, to our knowledge, to statistically evaluate the bias of face data in CycleGAN. In the above works, Kenfack *et al.* focused on classic GANs, and Jain *et al.* also primarily focused on traditional GANs (DCGAN and ProGAN). In our paper, we focus on CycleGAN, an image-to-image translation GAN model.

(2) We made a clear distinction between non-uniform and uniform training datasets, utilizing a skewed distribution of face-shots from the executive and board team members of the top-100 technology companies (Apple, Microsoft, Alphabet (Google), etc.).

(3) Our paper performs a case study on a real-world application for CycleGAN, namely, for Snapchat's gender-swap and gender-blending filter. By extending Jain *et al.* [41]'s study on Snapchat filters lightening skin color through including a larger representative dataset, we analyze how this filter reacts to the sensitive features that this paper discusses.

4 Implementation

In this section, we present both the training and testing datasets we used and how we collected and processed that data. We then discuss the model we implemented and our training environment on Google Colab.

4.1 Dataset Collection and Processing

In order to test our hypothesis that generator G will collapse to the modes in the majority group, we train a CycleGAN to translate the faces of non-tech-company members to look like tech companies' executive and board members. Thus, our output (target) domain contains a tech company member dataset (described below), and our input domain is a representative dataset (described further below).

We constructed a dataset of headshots of the executive and board members of the top-100 technology companies (Apple, Microsoft, Alphabet, etc.) [17]. We located the top-100 largest technology companies by market cap that had leadership (executive and board) directories with images that were available for public access. We used this choice of data because the diversity in high tech companies is publicly known to be skewed with regard to gender and race, according to statistics from [42][43]. As a result, it is an appropriate population to use for our study because we aim to test the aggravation of bias in an already-biased training set in GAN-based data generation. We collected a total of 1,136 tech company executive and board member headshots for our study. Using a Photograph action, we then cropped each of these images to only include the face, with about two inches of their neck included in the cropped image, and one inch above their head, and then resized all the images to 256 x 256-pixels. After cropping and resizing our images, we excluded images of low quality that were either blurry or had dimensions of less than 150 x 150 pixels. We included extra space around the face because our CycleGAN algorithm will resize the image to 286 x 286-pixels, and then randomly crop that image back to 256 x 256 pixels [31]. To obtain clearer results, we also centered the faces with the nose being located near the midpoint of the photo. We randomly split the 1,136 into two groups: the training dataset and testing dataset. We had a total of 1,036 images in our training dataset and 100 images in our testing dataset.

We then created our second collection of face images for training, which is the input domain and acts as a representative dataset for us to transform into tech company member faces through training on the first collection, modifying the faces to synthetically look more like technology company members. For this second collection of images, we used the Chicago Face Database [18]. The Chicago Face Database includes standardized and high-resolution images of female and male faces between 17-65 years of age, from a variety of different ethnicities. This second collection acts as a representative dataset for us to transform into tech company member faces through training on the first collection, modifying the faces to synthetically look more like technology company members, like a "filter". We used images from the Chicago Face Database because it is a significantly more representative sample than our tech company members' headshot dataset. Of this database, we used the main "CFD" set which contains photographs of 597 unique individuals, who self-identified as White, Latinx, Black, and Asian male and females from the United States. These images were pre-annotated with the models' self-identified ethnicity and gender. To obtain the most accurate results with regard to the tech company members dataset, we only chose to include the images of the models with neutral, happy (open mouth), or happy (closed mouth) facial expressions. We ended up with a total of 1,156 images from the Chicago Face Database. We randomly split these images to make up our training and testing datasets, which had 1,056 and 100 images, respectively. Our Chicago Face testing dataset contained 25 images from each of the four categories being tested (male white, male non-white, female white, and female non-white) for a total of 100 images.

4.2 Model Training

In order to explore the diversity of our GAN-generated images compared to the original images, we test the performance using a Pytorch CycleGAN implementation [31]. See Appendix for the GitHub source code. To obtain the synthetic images while accounting for variance in model training, we generated our test set images after training the CycleGAN for 200 epochs.

The computing environment is hosted in Google Colab. We initially tried to train the model using GPU from Google Cloud or Amazon Web Services (AWS), but eventually found that the managed environment in Google Colab is cost efficient and easier to operate. During training, we tuned the batch size parameter to leverage the high-power GPU that Google Colab offers. This

way, we controlled our training time within 18 hours. Google Colab is able to connect to Google Drive so all the training data, temporary results, and trained models could be directly saved to Google Drive. This allowed us to pause and resume training with ease. We linked our study's GitHub repository in the Appendix, which includes our trained model and CycleGAN-generated sample testing images.

Because CycleGANs are focused on image-to-image translation and are conditioned on the input domain images, unlike the traditional GAN variants, DCGAN and ProGAN [41], our first intuition was that they would be more immune to mode collapse that propagates biases in the original training set. However, it is known that even these image-to-image conditional GAN variants are susceptible to mode collapse [44]. Our research studies how CycleGANs react to sensitive social features like gender and race, which are still open questions.

5 Experimentation and Evaluation

We first conduct a demographic comparison experiment on our original and generated datasets by asking human volunteers to complete several study tasks provided in a survey. Then, we analyze the survey data results and the facial results obtained from our CycleGAN transformation. Finally, we conduct a case study on a real-world application of GAN.

5.1 Image-Translation Demographic Comparison

We posted a survey on the Nextdoor social media platform, and asked human volunteers from our neighborhoods to annotate face images from both the original dataset and the GAN-generated dataset along the latent dimensions of ethnicity and gender to assess the data. We explain our procedure below. To clarify, latent features are the unobservable variables that we must infer from other directly observable variables. In our survey, we asked each volunteer to perform four study tasks (T1a, T1b, T2a, and T2b). This procedure is similar to Jain *et al.* [41]'s procedure with DCGAN and ProGAN variants. The two datasets that we included to be annotated in this experiment are (1) the 100 randomly-selected technology company members testing dataset, DS_1, and (2) the 100 GAN-generated images, DS_2, that were translated to look like the technology company members. Each participant performed the following four tasks:

Task 1 (a/b): Subjects were asked to choose the most appropriate selection for an image that was from (T1a) DS_1 and (T1b) DS_2, from the following options: (1) face's skin tone is white (i.e pale), (2) face's skin tone is non-white, and (3) difficult to decide.

Task 2 (a/b): Subjects were asked to choose the most appropriate selection for an image that was from (T2a) DS_1 and (T2b) DS_2, from the following options: (1) face looks to be female, (2) face looks to be male, and (3) difficult to decide.

In total, we had 26 volunteers participate and fill out our survey form. Each subject was given 204 total images to annotate: 100 from the tech company data DS_1, 100 from the GAN-generated data DS_2, and 4 high-quality control images with known labels for skin color and gender. The observed results from the control images allowed us to eliminate 4 unreliable data points. In total, we had 22 valid data points from our volunteers' annotations for DS_1 and DS_2. To categorize each image as belonging to a class (female or male, and white or non-white) from the results, we used majority-voting. We also included the demographic information of the original 100 randomly-selected Chicago Database testing photos as a control, which is known to have equal distributions of male, female, white, and non-white individuals, and we did not need to conduct human-tasked annotation since these images were pre-annotated. We created this testing dataset to be representative with 25 randomly-selected images from each category: male white, male non-white, female-white, and female non-white.

5.2 Results

5.2.1 Human Study Tasks: Bias Aggravation

Our results from the human study tasks are presented in Figure6. Our representative testing dataset originally had an equal distribution in each of the four categories presented in the figure.

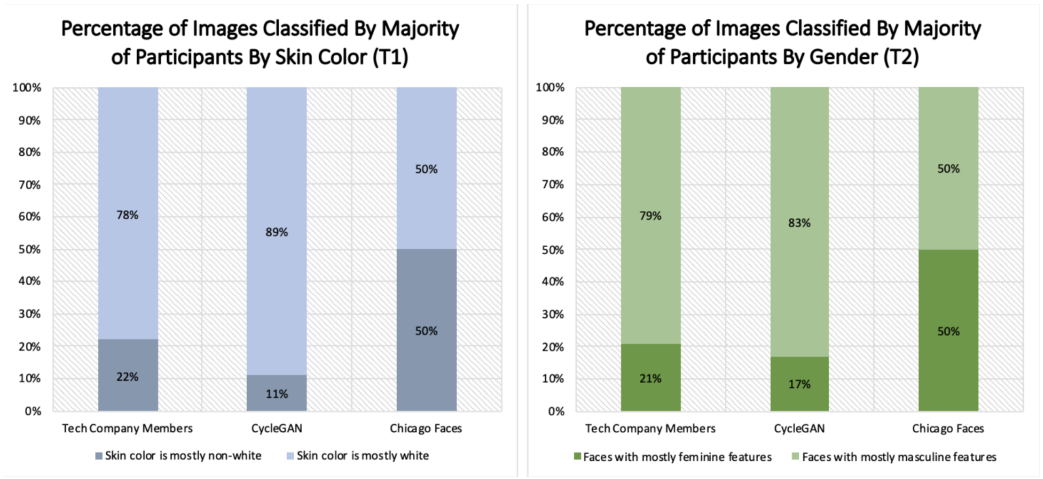

Figure 6: Distribution of human classifications on skin color (left) and gender (right). The left rows show the distribution of DS_1 and the middle rows show the distribution of DS_2. The right rows act as a control group and show the distribution of the representative Chicago Face testing dataset of 100 images.

With regard to female faces and male faces, our original dataset was 50/50 distributed. The same applied for our second graph which analyzed skin color being mostly non-white or white. We initially thought the CycleGAN we used in this study would propagate bias, but be less susceptible to further exacerbating the bias than other GANs, like concluded for DCGAN and ProGAN [41]. This is because, for instance, in our experiment where female faces are under-represented and gender is a latent feature, a GAN would have to perform the more challenging task of actively converting a female face to a male one. However, after applying our trained model on our representative testing dataset, according to the volunteers' average task results from above, (1) a majority of the non-white male and female faces in the representative sample were translated and classified as faces with mostly white skin color and (2) many of the female faces, both non-white and white, in the representative sample were translated and classified as faces that have mostly masculine features. Our results showing how CycleGAN transformed the input data are included in Figure6.

We plot the results for T1a and T1b in (Figure6, left) and found that CycleGAN penalized the original 22% of faces with mostly non-white skin color in DS_1, reducing this percentage to 11% in DS_2. We analyzed the statistical significance through a one-tailed two-proportion z-test, obtaining a z-score of 2.096 and a p-score of 0.0361. For a significance level of 0.05, the test outcome is statistically significant as 0.0361 is less than 0.05. Thus, we are also able to verify the exacerbation of bias along the latent dimension of ethnicity for CycleGAN. We observed that the synthetic face data generated by CycleGAN not only perpetuates, but aggravates biases against minority groups. Secondly, we plot the results for T2a and T2b in (Figure6, right) and found that CycleGAN also penalized the original 21% of faces with mostly feminine features in our tech members dataset DS_1, reducing this percentage to 17% in the GAN-generated dataset DS_2. This shows that CycleGAN propagates bias along the latent dimension of gender. With a z-score of 0.721 and a p-score of 0.471, we cannot prove statistical significance like we did for the latent dimension of ethnicity (skin color) in terms of the CycleGAN further aggravating bias to be more than the bias present in the original technology company members training dataset. We predict that this was due to CycleGAN not learning the transformation of certain features, namely hair, that likely affected the volunteers' labeling of these images, for both gender and skin tone, as they were able to catch some inconsistencies. We note that in our survey, we asked participants to analyze the skin tone (Task 1) based on lightness, not ethnicity. This was significant because for many of the images, the hair was a giveaway to the original ethnicity of the individual.

5.2.2 CycleGAN-Generated Transformation

In Figure7, we show examples of before and after transformations on faces that are representative of the minority classes (non-white, female) in the technology company members training dataset. It is interesting to note that the CycleGAN learned to add glasses and smiling facial expressions to the input image. This is because our model was trained on the technology company members input data, where many wore glasses and showcased a smiling expression for their headshot photos, whereas the individuals in the Chicago Face database did not wear glasses and a majority also had neutral facial expressions. However, not all of the modifications learned by the CycleGAN are socially harmless, as the model also translated masculine features onto the faces of women and lightened the skin color of non-white individuals. Of the 50 non-white men and women in our representative testing dataset of 100 images, we found that after translating them with our trained model, 39 of them were transformed and classified as white men or women. In addition, of the 50 women (both white and non-white) in our representative testing dataset of 100 images, we found that after translation, 33 of them were transformed and classified as faces with mostly masculine features.

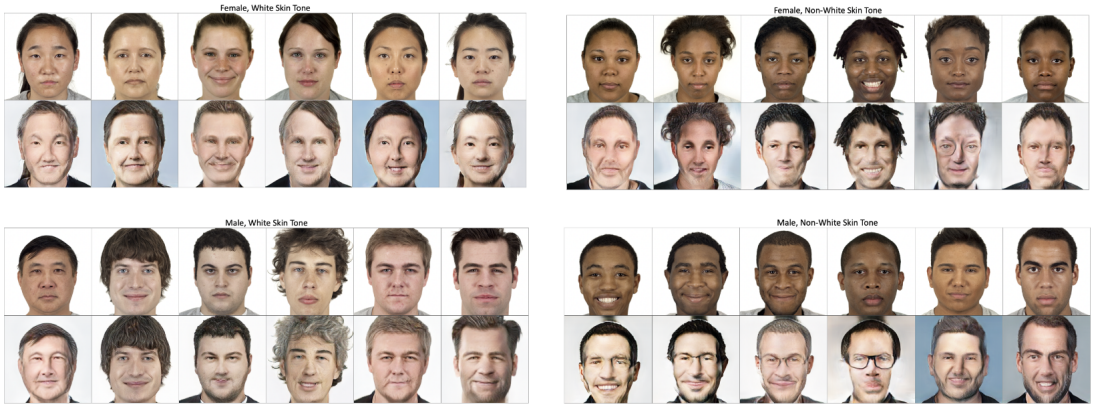

Figure 7: Examples of original face images and CycleGAN-transformed images to look like members of technology companies, for white skin-toned women (top left), non-white skin-toned women (top right), white skin-toned men (bottom left), and non-white skin-toned men (bottom right). For each of these four categories, the top row shows the original images and the bottom row shows the generated images.

Though it is expected that GAN could perpetuate biases not only along socially harmful dimensions, but along all arbitrary dimensions where there is a skew in the training dataset, this type of harmless bias is not the focus of our research. It is known that machine learning algorithms are designed to find patterns in the training data, but this can become problematic with respect to social features when models exacerbate biases for minority populations who have undergone systematic discrimination and disadvantage. The next section of our paper is a case-study where such GAN models are having adverse real-world implications.

5.3 Case Study: Snapchat "My Twin" Filter

Our experiment above served as an example of CycleGANs propagating bias, but the implications of mode collapse in GANs, specifically CycleGAN, can be seen in real-world applications, leading to bias-exacerbation. Snapchat, a popular image messaging platform, has recently started using image-to-image translation through GANs for several of their filters. This includes CycleGAN for their "My Twin" gender-swapping filter, our focus filter of this experiment, according to several sources [19] [20] [21]. We hypothesize that since this filter is presumably known to utilize the CycleGAN architecture, it may exacerbate biases that were present in the training data.

In order to evaluate how one vector, skin color, was altered between original and filtered images, we replicate a study presented in [41], using a different dataset: select individuals from

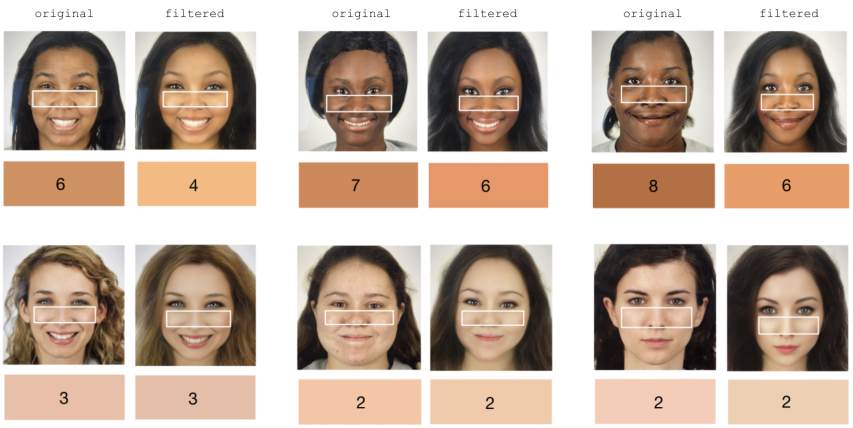

Figure 8: Faces of women of color (top row) and women of white skin tone (bottom row) before and after using Snapchat's female "My Twin" gender-swapping lens. The portions used for the skin-color mapping analysis are boxed in white on each image. The average pixel value color is shown below each image, which is a number corresponding to a lightness group on the L'Oréal skin tone chart [45].

the Chicago Face Database we used in our above experiment. We cropped a rectangular portion of the face that is below the eyes and above the tip of the nose. We then determined the average pixel value of this section, and use L2-norm distance to map this RGB vector to a shade on the L'Oréal skin tone chart [45], which we included in the Appendix of this paper. We decided to disregard skin warmth (A, B, C, etc.), and only took into account skin lightness. This is because the Snapchat filter tended to cause the skin tone to be warmer, ridding the face image of some of its initial redness. Before and after samples of our results are depicted in Figure8. We found that the gender-swapping filter lightened non-white faces by an average of 1.5 shades, and it lightened faces by one or two shades consistently for nine out of the ten non-white faces we tested. For white faces, on the other hand, the filter did not produce a significant effect. We found that it lightened three faces by one shade, darkened two faces by one shade, and didn't affect the other five faces. A more in-depth and rigorous research study should be performed to make further claims, but a potential cause for the lightening in skin tone of women of color is that the CycleGAN used by the filter undergoes mode collapse that collapses all inputs in a portion of the image to output colors that are lighter. Our case study provides initial evidence of the possibility that Snapchat filters like the "My Twin" filter used in this study lightens skin tone in women of color due to mode collapse.

6 Discussion: Real-World Applications

The bias-exacerbation implications of GANs in real-world applications can be severe. This is exemplified in our Snapchat case study, in which we showed that a CycleGAN-based gender-swapping filter lightened skin color only when used on women of color, and not on women of white skin tones. Since ideally the filter should not affect skin tone, this case study shows potential bias in the GAN model used to translate the faces based on the filter. However, in the scope of this paper, we have not performed a comprehensive study on this; thus, these claims are merely starting points that open up a research problem regarding bias in facial transformation filters, which are recently gaining popularity as the influence of social media thrives [46].

In downstream tasks, the consequences of using GAN-generated biased facial datasets can be adverse. Machine learning models used on facial data are currently widespread in important decision-making applications, including healthcare [12], employment [13], criminal justice [14], education [15], and security developments such as deep-fake detection [16]. The primary ethical problem with machine learning-based technology is that they tend to be used on populations to which they exhibit the most bias. Additionally, for individuals in marginalized communities, the

errors and misrepresentation caused by this bias is more costly than for other groups [47]. As a result, it is critical that from an ethical standpoint, GAN practitioners for data augmentation ensure that both training datasets and trained models are representative and diverse regarding sensitive features, such as race, ethnicity, and gender. We recommend GAN practitioners to analyze the distribution in training datasets and compare it to the distribution in generated datasets. It is important that we prevent these models from further under-representing populations that are already in the minority.

In criminal justice, there are automated, machine learning-powered tools aimed at predicting recidivism risk in a system that disproportionately punishes Latino and Black individuals. The risk assessment software that is applied in state criminal justice systems takes over 137 features as input, excluding race as a factor. Unfortunately, this risk assessment classification software is biased against the Black population, as it is found to disproportionately classify Black defendants as high or medium risk for recidivism [48]. A similar bias is seen in employment, where automated technology predicts candidate performance in industries that are already largely dominated by males, leading to the potential for further bias propagation. Furthermore, Amazon designed a hiring system in 2018 that was discovered to discriminate against female candidates as it penalizes résumés that included taking part in women's organizations [49]. This hiring system worked by collecting résumés from candidates over the span of ten years, yet it didn't take gender as an input feature. Another example of GANs' usage outside the range of classification is with PULSE [50], a self-supervised face "upscaler" and "depixelizer." When used to upsample photos of non-white faces, it was found to produce depixelized images with Caucasian features. In [50], the researchers performed studies that were able to conclude that these biases in the PULSE image upscaler drew directly from the biased performance of the GAN model from which it receives synthetic augmented data.

The primary takeaway from this discussion on real-world implications is that the GAN-bias challenge spans wider than merely the dataset used. It has been proven that GANs do not solely propagate the distribution patterns in the original training dataset, but amplify them. The question also expands to how society can regulate the real-world applications for which these systems are used.

7 Conclusion and Future Work

GANs have been shown through several works to lessen diversity in distributions when compared to the datasets they are trained on due to mode collapse. However, the implications of mode collapse are unclear in situations where the training dataset is biased and skewed toward certain feature values, such as men and white skin tones, along latent features, in this case gender and ethnicity (skin tone). To research this, we used a popular image-to-image translation GAN, CycleGAN, to stylize faces from a representational dataset to look like the faces of members of technology companies. To our knowledge, we are the first to empirically analyze how CycleGANs trained on skewed datasets toward male and white faces both perpetuate and aggravate social biases along the dimensions of gender and skin color when synthetically generating a new dataset. In our experiment, we found that there was a significant under-representation of non-white skin tones and feminine facial features in the generated dataset, caused by mode collapse on a majority latent mode of the training dataset. We then performed a case study on Snapchat's gender-swapping "My Twin" filter, demonstrating potential real-world implications of bias from mode collapse in CycleGANs. This work cautions the use of GAN-based data augmentation methods to address real-world challenges stemming from data limitation in downstream tasks. As the use of GANs expands with the increase in deep learning models used in various fields, there is a growing sense of security that GANs have the ability to elegantly generate unique novel data. In actuality, however, the generated data may be under-representing some critical features of the real-world data.

As future work, we aim to expand our study to analyze other features beyond skin tone and gender. In our current study, we evaluated the latent dimension of ethnicity based off skin tone, placing a majority of east Asian Americans in the "white" skin tone category, and a majority of Latinx people in the "non-white" skin tone category. We intend to distinguish between faces of

different ethnicities (i.e Asian, Caucasian, Black, Latino, etc.) through performing a study that evaluates whether CycleGAN collapses along these specific modes. We also aim to implement a larger dataset with double the amount of images in both the training and testing dataset to obtain more precise results. Additionally, we open up a more in-depth study on the implications we identified in Snapchat's "My Twin" filter, through an expansion including more participants and studying the GAN architecture behind the filter.

A Appendix

GitHub Repository: https://github.com/fionapeng16/CycleGAN-facial-bias

Pytorch CycleGAN implementation source code: [31]: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

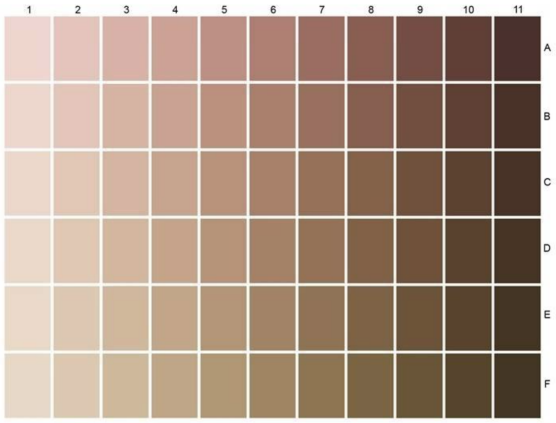

Figure 9: L'Oréal Research: "A New Geography of Skin Color" chart [45].

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Jun. 2014, arXiv:1406.2661 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1406.2661

[2] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "CT-Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, vol. 11071, pp. 732–740, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00934-2_81

[3] X. Wang, K. Wang, and S. Lian, "A Survey on Face Data Augmentation," *Neural Computing and Applications*, vol. 32, no. 19, pp. 15 503–15 531, Oct. 2020, arXiv:1904.11685 [cs]. [Online]. Available: http://arxiv.org/abs/1904.11685

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," Aug. 2020, arXiv:1703.10593 [cs]. [Online]. Available: http://arxiv.org/abs/1703.10593

[5] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," Nov. 2014, arXiv:1411.1784 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1411.1784

[6] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li, "Mode Regularized Generative Adversarial Networks," Mar. 2017, arXiv:1612.02136 [cs]. [Online]. Available: http://arxiv.org/abs/1612.02136

[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," Jan. 2016, arXiv:1511.06434 [cs]. [Online]. Available: http://arxiv.org/abs/1511.06434

[8] (2022) Bias and generalization in deep generative models: An empirical study. [Online]. Available: https://proceedings.neurips.cc/paper/2018/hash/5317b6799188715d5e00a638a4278901-Abstract.html

[9] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-Guided Face Generation Using Conditional CycleGAN," Nov. 2018, arXiv:1705.09966 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1705.09966

[10] (2022) Summitkwan/transparentlatentgan: Use supervised learning to illuminate the latent space of gan for controlled generation and edit. [Online]. Available: https://github.com/SummitKwan/transparent_latent_gan

[11] S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang, "Generalization and equilibrium in generative adversarial nets (GANs)," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 224–232. [Online]. Available: https://proceedings.mlr.press/v70/arora17a.html

[12] (2022) Can your face reveal how long you'll live? new technology may provide the answer. - the washington post. [Online]. Available: https://www.washingtonpost.com/national/health-science/can-your-face-reveal-how-long-youll-live-new-technology-may-provide-the-answer/2014/07/02/640bacb4-f748-11e3-a606-946fd632f9f1_story.html

[13] (2019) Ai used for first time in job interviews in uk to find best applicants. [Online]. Available: https://www.telegraph.co.uk/news/2019/09/27/ai-facial-recognition-used-first-time-job-interviews-uk-find/

[14] (2019) Amazon's facial-recognition ai is supercharging police in oregon. but what if rekognition gets it wrong? - the washington post. [Online]. Available: https://www.washingtonpost.com/technology/2019/04/30/amazons-facial-recognition-technology-is-supercharging-local-police/

[15] T. M. Harmeet Kaur. (2019) A new york school district is bringing in facial recognition software. rights groups say it could spell trouble for students — cnn. [Online]. Available: https://www.cnn.com/2019/05/30/us/ny-school-facial-recognition-trnd/index.html

[16] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The DeepFake Detection Challenge (DFDC) Dataset," 2020, publisher: arXiv Version Number: 4. [Online]. Available: https://arxiv.org/abs/2006.07397

[17] (2022) Largest tech companies by market cap. [Online]. Available: https://companiesmarketcap.com/tech/largest-tech-companies-by-market-cap/

[18] D. S. Ma, J. Correll, and B. Wittenbrink, "The Chicago face database: A free stimulus set of faces and norming data," *Behav Res*, vol. 47, no. 4, pp. 1122–1135, Dec. 2015. [Online]. Available: http://link.springer.com/10.3758/s13428-014-0532-5

[19] (2022) Gender swap and cyclegan in tensorflow 2.0 — flipboard. [Online]. Available: https://flipboard.com/article/gender-swap-and-cyclegan-in-tensorflow-2-0/a--l2JlSV6R6CqZqrSYlHz2g%3Aa%3A2892075988-4be10a8b1c%2Ftowardsdatascience.com

[20] (2019) The dark implications of facial swap filter technology - paper. [Online]. Available: https://www.papermag.com/snapchat-gender-swapping-filter-2638765039.html?rebelltitem=15#rebelltitem15

[21] (2022) (1) [d] is the new snapchat gender filter gan-based? : Machinelearning. [Online]. Available: https://www.reddit.com/r/MachineLearning/comments/bo4orw/d_is_the_new_snapchat_gender_filter_ganbased/

[22] Z. Wang, Q. She, and T. E. Ward, "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy," Dec. 2020, arXiv:1906.01529 [cs]. [Online]. Available: http://arxiv.org/abs/1906.01529

[23] I. Goodfellow, "NIPS 2016 Tutorial: Generative Adversarial Networks," Apr. 2017, arXiv:1701.00160 [cs]. [Online]. Available: http://arxiv.org/abs/1701.00160

[24] L. O'Gorman and R. Kasturi, *Executive briefing: document image analysis*. Los Alamitos, Calif: IEEE Computer Society Press, 1997.

[25] D. A. Teich. (2019) Synthetic data is a tool for improving training and accuracy of deep learning systems. [Online]. Available: https://www.forbes.com/sites/davidteich/2019/05/28/synthetic-data-is-a-tool-for-improving-training-of-deep-learning-systems/?sh=eab6b8b7b7f1

[26] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Nov. 2018, arXiv:1611.07004 [cs]. [Online]. Available: http://arxiv.org/abs/1611.07004

[27] Y. Hiasa, Y. Otake, M. Takao, T. Matsuoka, K. Takashima, A. Carass, J. L. Prince, N. Sugano, and Y. Sato, "Cross-Modality Image Synthesis from Unpaired Data Using CycleGAN: Effects of Gradient Consistency Loss and Training Data Size," in *Simulation and Synthesis in Medical Imaging*, A. Gooya, O. Goksel, I. Oguz, and N. Burgos, Eds. Cham: Springer International Publishing, 2018, vol. 11037, pp. 31–41, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00536-8_4

[28] J. de Curtó and R. Duvall, "Cycle-consistent Generative Adversarial Networks for Neural Style Transfer using data from Chang'E-4," Nov. 2020, arXiv:2011.11627 [cs]. [Online]. Available: http://arxiv.org/abs/2011.11627

[29] A. Jarda. (2020) A gentle introduction to cycle consistent adversarial networks — by aamir jarda — towards data science. [Online]. Available: https://towardsdatascience.com/a-gentle-introduction-to-cycle-consistent-adversarial-networks-6731c8424a87

[30] (2021) Cyclegan for image to image translation. [Online]. Available: https://blog.jaysinha.me/train-your-first-cyclegan-for-image-to-image-translation/

[31] (2022) junyanz/pytorch-cyclegan-and-pix2pix: Image-to-image translation in pytorch. [Online]. Available: https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

[32] Y. Mirsky, T. Mahler, I. Shelef, and Y. Elovici, "CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning," Jun. 2019, arXiv:1901.03597 [cs]. [Online]. Available: http://arxiv.org/abs/1901.03597

[33] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical Image Synthesis with Context-Aware Generative Adversarial Networks," in *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, vol. 10435, pp. 417–425, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-66179-7_48

[34] (2019) Using cyclegan for age conversion — paperspace blog. [Online]. Available: https://blog.paperspace.com/use-cyclegan-age-conversion-keras-python/

[35] X. Zhu, Y. Liu, J. Li, T. Wan, and Z. Qin, "Emotion Classification with Data Augmentation Using Generative Adversarial Networks," in *Advances in Knowledge Discovery and Data Mining*, D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, Eds. Cham: Springer International Publishing, 2018, vol. 10939, pp. 349–360, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-319-93040-4_28

[36] P. J. Kenfack, K. Sabbagh, A. R. Rivera, and A. Khan, "RepFair-GAN: Mitigating Representation Bias in GANs Using Gradient Clipping," Jul. 2022, arXiv:2207.10653 [cs]. [Online]. Available: http://arxiv.org/abs/2207.10653

[37] P. Grnarova, K. Y. Levy, A. Lucchi, N. Perraudin, I. Goodfellow, T. Hofmann, and A. Krause, "A domain agnostic measure for monitoring and evaluating GANs," Jul. 2020, arXiv:1811.05512 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1811.05512

[38] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, "Unrolled Generative Adversarial Networks," 2016, publisher: arXiv Version Number: 4. [Online]. Available: https://arxiv.org/abs/1611.02163

[39] D. Mishra, P. A. P., A. Jayendran, V. Srivastava, and S. Chaudhury, "Mode matching in GANs through latent space learning and inversion," Mar. 2019, arXiv:1811.03692 [cs]. [Online]. Available: http://arxiv.org/abs/1811.03692

[40] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, publisher: arXiv Version Number: 3. [Online]. Available: https://arxiv.org/abs/1701.07875

[41] N. Jain, A. Olmo, S. Sengupta, L. Manikonda, and S. Kambhampati, "Imperfect ImaGANation: Implications of GANs Exacerbating Biases on Facial Data Augmentation and Snapchat Selfie Lenses," Jun. 2021, arXiv:2001.09528 [cs, eess, stat]. [Online]. Available: http://arxiv.org/abs/2001.09528

[42] (2022) Diversity in high tech statistics [2022] – zippia. [Online]. Available: https://www.zippia.com/advice/diversity-in-high-tech-statistics/

[43] (2019) Diversity in tech by the numbers: Age, race, gender recruiting innovation. [Online]. Available: https://recruitinginnovation.com/diversity-in-tech/

[44] S. Ma, J. Fu, C. W. Chen, and T. Mei, "DA-GAN: Instance-level Image Translation by Deep Attention Generative Adversarial Networks (with Supplementary Materials)," Feb. 2018, arXiv:1802.06454 [cs]. [Online]. Available: http://arxiv.org/abs/1802.06454

[45] (2022) Expert in skin and hair types around the world. [Online]. Available: https://www.loreal.com/en/articles/science-and-technology/expert-inskin/

[46] D. Chaffey. (2022) Global social media statistics research summary 2022 [june 2022]. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[47] T. Gebru, "Oxford Handbook on AI Ethics Book Chapter on Race and Gender," Aug. 2019, arXiv:1908.06165 [cs]. [Online]. Available: http://arxiv.org/abs/1908.06165

[48] (2016) Machine bias — propublica. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[49] J. Dastin. (2018) Amazon scraps secret ai recruiting tool that showed bias against women — reuters. [Online]. Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[50] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models," 2020, publisher: arXiv Version Number: 3. [Online]. Available: https://arxiv.org/abs/2003.03808

[51] *2018 IEEE High Performance Extreme Computing Conference (HPEC): the Westin-Hotel Waltham-Boston, 70 Third Avenue, Waltham, Massachusetts USA, 25-27 September 2018*. Piscataway, New Jersey: IEEE, 2018, oCLC: 1078908223.

[52] V. Sandfort, K. Yan, P. J. Pickhardt, and R. M. Summers, "Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks," *Scientific Reports*, vol. 9, no. 1, p. 16884, Dec. 2019. [Online]. Available: http://www.nature.com/articles/s41598-019-52737-x

[53] Z. Zhou and C. Firestone, "Humans can decipher adversarial images," *Nature Communications*, vol. 10, no. 1, p. 1334, Mar. 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-019-08931-6

[54] Q. Zhou, M. Zuley, Y. Guo, L. Yang, B. Nair, A. Vargo, S. Ghannam, D. Arefan, and S. Wu, "A machine and human reader study on AI diagnosis model safety under attacks of adversarial images," *Nature Communications*, vol. 12, no. 1, p. 7281, Dec. 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-27577-x