UNIVERSITY OF CALIFORNIA SAN DIEGO

Debiasing Image Generative Models

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Md Mehrab Tanjim

Committee in charge:

Professor Garrison W. Cottrell, Chair
Professor Taylor Berg-Kirkparick
Professor Sanjoy Dasgupta
Professor Virginia de Sa
Professor David Kriegman

2023

The Dissertation of Md Mehrab Tanjim is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my advisor, Prof. Gary Cottrell, for his guidance and support throughout my thesis journey. He has been an invaluable source of knowledge, inspiration, and encouragement. His expertise in computer vision & generative models and his invaluable feedback on my research have been crucial in shaping the direction and quality of my work. I am deeply grateful for his patience and understanding during the many challenges and setbacks that I faced. He always went above and beyond to provide me with the resources and support I needed to succeed. Thank you, Prof. Gary Cottrell, from the bottom of my heart.

I would like to express my gratitude to my collaborators from Adobe Research, Ritwik Sinha and Krishna Kumar Singh, for their invaluable support and company throughout our various projects and papers. I also extend my thanks to other collaborators from Adobe Research, including Moumita Sinha, Kushal Kafle, Vishy Swaminathan, David Arbour, and Sridhar Mahadevan, for their contributions.

I am deeply appreciative of the guidance and mentorship provided by my committee member and several professors at UC San Diego: Prof. Taylor Kirkpatrick, Prof. Sanjoy Dasgupta, Prof. David Kriegman, Prof. Virginia de Sa, and Prof. Lawrence Saul. Working with these exceptional researchers has shaped my perspective and attitude, and I am extremely fortunate to have had the opportunity to learn from them.

Above all, I owe my deepest thanks to my family, especially my parents: Md Liakot Ali & Kamrun Naher, and my brother: Md Fahim Anjum, whose immense sacrifices and support have made everything possible. Finally, I would also like to thank my friends and seniors: Hammad Ayuubi, Hasan Imam, Farhana Rahman, Iftekhar Chowdhury, Iftikhar Ahmad Niaz, Ahmed Naguib, and Mohammed Alyaseen, for providing me with the essential support and social structure in a foreign country that allowed me to thrive academically.

Chapter 2, in full, is a reprint of the material as it appears in IEEE Winter Conference on Applications of Computer Vision (WACV), 2022. **Md Mehrab Tanjim**, Ritwik Sinha,

Krishna Kumar Singh, Sridhar Mahadevan, David Arbour, Moumita Sinha, Garrison W. Cottrell, "Generating and Controlling Diversity in Image Search". The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in British Machine Vision Conference (BMVC), 2022. **Md Mehrab Tanjim**, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell, "Debiasing Image-to-Image Translation Models". The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Responsible Computer Vision at European Conference on Computer Vision (RCV@ECCV), 2022. **Md Mehrab Tanjim**, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell, "Discovering and Mitigating Biases for CLIP-based Text-to-Image Generation". The dissertation author was the primary investigator and author of this paper.

VITA

2017	Bachelor of Science, Bangladesh University of Engineering and Technology
2017-2018	Research Assistant, Bangladesh University of Engineering and Technology
2019-2021	Teaching Assistant, University of California San Diego
2021	Master of Science, University of California San Diego
2018-2023	Research Assistant, University of California San Diego
2023	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell, "Debiasing Image-to-Image Translation Models", British Machine Vision Conference (BMVC), 2022

Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell, "Discovering and Mitigating Biases for CLIP-based Text-to-Image Generation", Responsible Computer Vision at European Conference on Computer Vision (RCV@ECCV), 2022

Md Mehrab Tanjim, Ritwik Sinha, Krishna Kumar Singh, Sridhar Mahadevan, David Arbour, Moumita Sinha, Garrison W. Cottrell, "Generating and Controlling Diversity in Image Search", IEEE Winter Conference on Applications of Computer Vision (WACV), 2022.

T.M. Tariq Adnan, **Md Mehrab Tanjim** and Muhammad Abdullah Adnan, "Fast, Scalable and GeoDistributed PCA for Big Data Analytics", Elsevier Journal on Information Systems, 2021

Md Mehrab Tanjim, Hammad A. Ayyubi, Garrison W. Cottrell, "DynamicRec: A Dynamic Convolutional Network for Next Item Recommendation", ACM International Conference on Information & Knowledge Management (CIKM), 2020

Md Mehrab Tanjim, Congzhe Su, Ethan Benjamin, Diane Hu, Liangjie Hong and Julian McAuley, "Attentive Sequential Models of Latent Intent for Next Item Recommendation", International World Wide Web Conference (WWW), 2020

Wang-cheng Kang, **Md Mehrab Tanjim**, Lucky Dhakad, Surender Kumar, Julian McAuley, "Dynamically Predicting Budgets and Purchase Intent with Neural Sequential Models", Context Aware Recommender Systems (CARS) Workshop, 2020.

Md Mehrab Tanjim and Muhammad Abdullah Adnan, "sSketch: A Scalable Sketching Technique for PCA in the Cloud", ACM International Web Search & Data Mining (WSDM) Conference, 2018

ABSTRACT OF THE DISSERTATION

Debiasing Image Generative Models

by

Md Mehrab Tanjim

Doctor of Philosophy in Computer Science

University of California San Diego, 2023

Professor Garrison W. Cottrell, Chair

Generative models have become increasingly popular in various domains to solve challenging tasks, including image generation, dialogue generation, and story generation. Unlike discriminative models, they can learn the underlying probability distribution of data and generate new examples. In particular, image generative models have gained significant attention due to their remarkable ability to produce images of unparalleled quality. However, while there has been a lot of attention to biases in discriminative models, bias in generative models has received little attention. The presence of biases in generative models, particularly related to race and gender, can have significant consequences in downstream applications. Therefore, efforts to address this issue are essential to promote fair and ethical use of generative models in various

domains. To achieve this goal, this dissertation presents a comprehensive study of debiasing image generative models by incorporating diversity and fairness constraints into the training process.

In this dissertation, we investigate three different approaches to debiasing image generative models. In the first approach, a new task of high-fidelity image generation conditioned on multiple attributes from imbalanced datasets is proposed. This task poses new challenges for state-of-the-art GANs models, and a new training framework is proposed to address these challenges. The second approach investigates bias in image-to-image translation models and proposes debiasing using contrastive learning. Finally, the study highlights the prevalence of bias in large-pretrained models like CLIP and its impact on text-to-image generative models. Identity preserving losses are proposed to rectify the problem without retraining the pretrained model. In all of these approaches, we evaluate the impact of debiasing on image generation and the effectiveness of existing methods in reducing biases in generated images. We show the proposed task and framework offer new avenues for further research in debiasing generative models. Overall, this dissertation contributes to the field of generative models by providing a comprehensive study of debiasing generative models and proposing a new task and framework for high-fidelity image generation.

Chapter 1

Introduction

Generative models have been recently gaining a lot of popularity in many domains to solve various interesting and challenging tasks such as dialogue generation [40], review synthesis [49], story generation [18], generating photorealistic images [32], super-resolution [39], etc. A generative model is different from a discriminative network in the sense that a generative model can learn the underlying probability distribution of the data (either implicitly or explicitly) to synthesize new examples. Especially in the domain of images, generative models have shown a lot of promise.

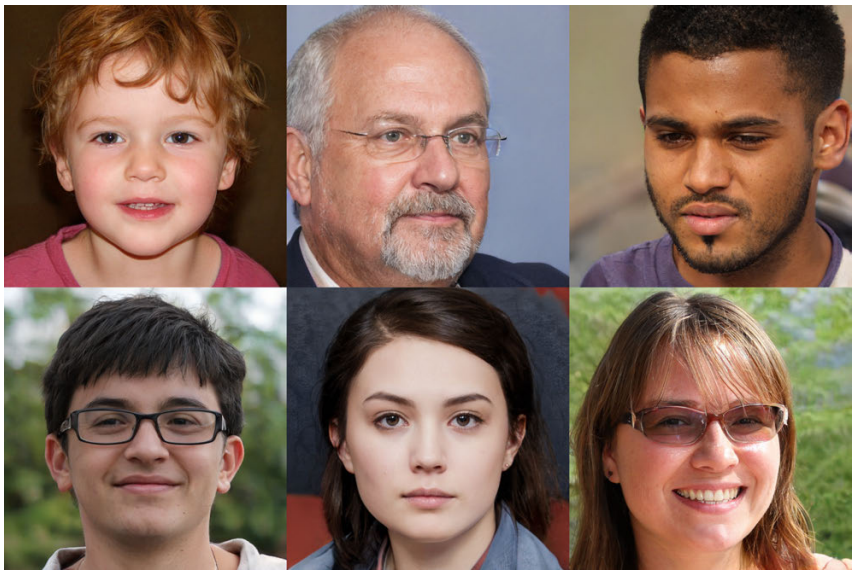


Figure 1.1. High-resolution photo realistic image generated by StyleGAN2 [33]. It is quite difficult, even for a human, to tell whether these persons are fake or not.

Earlier work on image generation was primarily focused on models with parametric specification of the probability distribution. The most noticeable work in this direction are PixelCNN [67] and Variational Autoencoders [38]. Generative Adversarial Networks or GANs [19], on the other hand, do not explicitly parameterize the data distribution. Instead of learning the parameters through maximum likelihood, GANs aim to sample new examples by learning an approximation to a target data distribution by training two components: generators and discriminators. The generators take random input and try to transform it to the data distribution while the discriminator tries to distinguish between generated samples and real examples. Although training GAN models is relatively difficult compared to other models, the quality of generated examples from GANs exceed any of the previous image generative models. Especially with the availability of GPU resources and a large amount of data, GANs are able to achieve unprecedented image quality, so much so it is now practically impossible for a human to discern the difference between real vs fake images. For example, Figure 1.1 shows photo-realistic high-resolution images generated by a state-of-the-art GAN model, StyleGAN2 [32]. As can be seen from the figure, it is quite challenging to tell whether these persons are real or not.

While much research has been dedicated to enhancing the quality of generated images, little attention has been paid to how biases in the training data or models can affect image results. Such biases often revolve around race and gender, which can significantly impact various downstream applications. As a result, debiasing generative models has become a critical area of research as it helps to ensure that these models are more equitable and unbiased in their outputs. This dissertation investigates different approaches to debiasing generative models and assesses their efficacy. A brief overview of these approaches is presented below. Later, we describe each of them in detail in their respective chapter.

1.1 Debiasing Image Search

In our society, generations of systemic biases have led to some professions being more common among certain genders and races. This bias is also reflected in image search on stock image repositories and search engines, e.g., a query like "male Asian administrative assistant" may produce limited results. The pursuit of a utopian world demands providing content users with an opportunity to present any profession with diverse racial and gender characteristics. The limited choice of existing content for certain combinations of profession, race, and gender presents a challenge to content providers. Current research dealing with bias in search mostly focuses on re-ranking algorithms. However, these methods cannot create new content or change the overall distribution of protected attributes in photos. To remedy these problems, we propose a new task of *high-fidelity* image generation conditioning on multiple attributes from *imbalanced* datasets. Our proposed task poses new sets of challenges for the state-of-the-art Generative Adversarial Networks (GANs). In this work, we also propose a new training framework to better address the challenges. We evaluate our framework rigorously on a real-world dataset and perform user studies that show our model is preferable to the alternatives. This work is described in Chapter 2.

1.2 Debiasing Image-to-Image Translation Models

Deep generative models have shown a lot of promise in various image-to-image translation tasks such as image enhancement and generating images from sketches. However, when all the classes are not equally represented in the training data, these algorithms can fail for underrepresented classes. For example, our experiments with the CelebA-HQ face dataset [30] reveal that this bias is prevalent for infrequent attributes, e.g., eyeglasses and baldness. Even when the input image clearly has eyeglasses, the image translation model is unable to create a face with them. To remedy this problem, we propose a data and model agnostic, general framework based on contrastive learning, re-sampling, and minority category supervision to

debias existing image translation networks for various image-to-image translation tasks such as super-resolution and sketch-to-image. Our experimental results from the real and synthetic datasets show that our framework outperforms the baselines both quantitatively and qualitatively. This work is described in Chapter 3.

1.3 Debiasing CLIP-based Text-to-Image Generative Models

Recently, CLIP (Contrastive Language-Image Pre-Training) [54] is gaining popularity in various downstream applications, such as zero-shot image classifiers, text-to-image synthesis, etc. However, despite being trained on a large dataset, the CLIP model suffers from biases against protected attributes such as gender and race. While earlier work focuses on the implication of such biases for image classification, to the best of our knowledge, no similar study has been done for CLIP-based generative tasks. In this work, we first reveal the queries for which the CLIP model biases the generated images in the text-to-image synthesis task. We also propose several ways to mitigate the biases without retraining CLIP or the underlying generative model. This work is described in Chapter 4.

Chapter 2

Generating and Controlling Diversity in Image Search

2.1 Introduction

Due to historic stereotypes that exist in our society, image search results can become biased for certain sets of queries. This problem is particularly extreme for certain professions, for example, Figure A.13 shows the top search results for the profession 'plumber' from Google Image search[1]. As we can see, most top results are of young white men; this is a reflection of societal stereotypes for this occupation. Similar types of results exist for other queries of various professions such as 'carpenter', 'machine operator', 'administrative assistant', 'cleaner', and so on, where the search results reveal biases in the gender, ethnicity, and age in the top results. Unsurprisingly, due to such societal bias, some combinations of race and gender may have few or no images in a content repository. For example, when we searched 'female black (or African American) machine operator' or 'male Asian administrative assistant', we did not find relevant images on Google Image search[2] In addition, in rare instances, particular combinations of gender and race can lead to individuals being portrayed inappropriately. We observed this behavior for search queries like 'female Asian plumber' or 'female Black (or African American) security guard.' This type of behavior is unwanted as it leads to dissatisfied consumers. This problem

[1]Search conducted in January, 2021 from California.

[2]When we first conducted this research. Search engines may have been updated since.

Figure 2.1. Top image results retrieved from Google Image search for the query 'plumber' reveal intrinsic biases.

affects both image search and stock platforms with paying customers.

In the presence of such paucity of content, current bias-mitigating re-ranking algorithms are not helpful because they seek to re-order existing images relevant to a query [28, 8], but cannot create new content nor increase the overall diversity within the results. For example, if there is only one picture of a 'male Asian administrative assistant', existing strategies will not help the user experience. Instead, imagine a machine that can generate photo-realistic high-resolution images for such queries. Such engines would tremendously enrich the user experience if end-consumers can access new content for any combination of attributes. Real images may not exist, or if they do, there might be only a few images with little or no variation, or in the worst case be inappropriate images. For such an application, generative models, in particular, Generative Adversarial Networks (or GANs) [19], have great potential because of their ability to produce photo-realistic images either unconditionally [32, 33] or conditionally [80, 41, 44, 42].

In light of these considerations, to address this bias and lack of diversity in image search, we propose a new task: generating *high-resolution* images controlling for *multiple* attributes, from *imbalanced* datasets. This task raises several new challenges. First, it is hard to define specifically what to visualize when creating new content for different occupations. A real image can be incredibly complex because of diverse backgrounds, various accessories, multiple people,

Figure 2.2. High-resolution images generated for the set of keywords from our proposed model 'Uniform+.'

and so on (which is apparent in Figure A.13). Therefore, directly collecting images from search results using different queries will not lead to an optimal and clean dataset for training GAN models. Second, for content to be consumable by the end-user, the generated images need to be available in high-resolution. Unfortunately, current state-of-the-art GAN models for high-quality (HQ) image generation, such as StyleGAN [32] or StyleGAN2 [33], learn image features without any supervision and do not allow explicit control over attributes. While we can augment these models with class-conditioning, trivial conditioning on attributes will not be sufficient for our task because the imbalance in the training dataset across multiple classes (such as race, gender and occupation) propagates to the generated images. Finally, we have observed that the automatic metrics to evaluate the quality of the generated images, such as Frechet Inception Distance (FID) [22] and classification accuracy, cannot sufficiently measure the image quality for our proposed task. To rectify these challenges, we make the following contributions:

- To explicitly control the image generation process, we first augment the state-of-the-art GAN model, StyleGAN2, with multi-class conditioning. To overcome the imbalance in the dataset, we compare two training procedures: weighted loss and over-sampling the minority class. Based on our finding from the comparison, we come up with a new training

procedure that combines over-sampling with image augmentation which can effectively handle the multiple-class imbalance. This training procedure is not specific to StyleGAN2 and thus can be applied to any generative model to combat bias in the dataset.

- As there is no existing dataset to train such models for debiasing image search results, we also build a new high-quality dataset for this task (which we call Stock-Occupation-HQ) and we describe the guidelines for the data collection, pre-processing, and annotation.

- Finally, we conduct both quantitative and qualitative evaluations to compare the performance of all models. For quantitative evaluation, we calculate the widely used metric FID [22] and classification accuracy (similar to [11]) which we call Attribute Matching Score. But our experimental results reveal a tradeoff between these two metrics and prove them insufficient to gauge the comparative quality of images. So, for qualitative evaluation, we perform user studies on Amazon Mechanical Turk (AMT) which show the strength of our proposed approach.

Generated images from our best performing model, Uniform+, are demonstrated in Figure 2.2 which show exciting results for combating bias in image search.

2.2 Related Work and Background

2.2.1 Bias in Image Search

To characterize the gender bias in image search results for a variety of occupations, the authors of [34] collected the top 400 image results for 96 occupations from Google images, and used human annotators to label them. They showed that the percentage of images of women in Google's 2014 results was 37%, and the fraction of gender anti-stereotypical images was only 22%, a number lower than expected. Moreover, they showed that sometimes images from gender minorities are portrayed unprofessionally. They call this the 'sexy carpenter' problem. A more recent study [8] shows that diversity in search has improved in the last five years, but not too

significantly. For example, they show that the percentage of female participants has risen to 45% in 2019, but the fraction of anti-stereotypical images has remained low (30% in Google 2019). To mitigate such bias in search results, current research mainly focuses on developing re-ranking algorithms that can show diversity in the top search results. For example, [28] propose a Fairness Maximal Marginal Relevance (FMMR) retrieval algorithm to reflect diversity in the top image search results. Similar work is explored in [8]. However, these methods can only mitigate bias in the top results by re-ranking if many diverse images relevant to the query exist. This may not hold for combinations of racial and gender attributes that are less common for a certain profession. When these images do not exist, or only a few of them do, these methods cannot diversify the overall search results. This suggests the need for a generative solution, where we can always generate new content for any mixture of attributes.

2.2.2 Attribute-to-Image Synthesis Models

In recent years, Generative Adversarial Networks or GANs [19] have become very popular in the domain of image generation. Originally, GANs were proposed to unconditionally generate images from random noise. To exert control over the generation process, GANs conditioned on class labels [48, 44, 73, 6] or text input have been proposed [80, 42, 41]. As these models allow the explicit control of generation conditional on attributes, we can potentially apply them to our proposed task. However, a common limitation of these models is their lack of ability to produce images at high-resolution, which is one of the requirements for platforms that provide content, like image search providers or stock image platforms.

2.2.3 High-Quality (HQ) Image Generation

For content platforms, the resolution of attribute-controlled generated images needs to be as high as possible (preferably 1024×1024). Generating such high-quality images, however, is significantly difficult because, at high resolution, it becomes easier for the discriminator to tell the fake images from real ones and training can easily become unstable. For example,

one of the class-conditioned generative models, BigGAN [6], can produce results at 512×512 pixel. Even at this smaller resolution (half of what is required), they show their models undergo training collapses. Additionally, results for BigGAN are shown in the balanced dataset setting (where each image belongs to only one class). Without class-conditioning, there exist only a handful of models that can generate images at such a high-resolution spectrum. For example, to stabilize the training process at high-resolution, a progressive GAN is proposed in [30], where they grow the resolution of both generator and discriminator progressively, from 4×4 to 1024×1024. However, a key problem of that architecture is feature entanglement: it represents faces holistically, which makes it difficult to modify eyes, for example, independently from the rest of the face. StyleGAN [32] and its improved version StyleGAN2 [33] both combat this entanglement problem by introducing a mapping network and adaptive instance normalization (AdaIN) [23] into the progressive GAN.

However, both StyleGAN and StyleGAN2 learn disentangled representations from images without any supervision and do not allow explicit control over attributes, which is crucial for our task. Furthermore, they do not have any built-in mechanism that allows them to train under class imbalance, where only a few examples exist for certain combinations of attributes. In this work, we overcome these new challenges in our proposed task.

2.3 Our Approach

Our objective for this section is to propose models suitable for our new task of generating HQ images for a rare combination of attributes to mitigate bias.

2.3.1 Base Network Selection

For choosing a base network, our priority is to make sure the synthesis model can generate high-quality images. Specifically, the model needs to generate faces in great details because they have to reflect the sensitive attributes clearly, such as race and gender. For these reasons, we have found in our early experiments that the current attribute-controlled image-to-image

translation systems such as STGAN[44] and text-to-image synthesis generative models such as DMGAN[80], CPGAN[42], Obj-GAN[41] were not a good fit , as the quality of images degraded at high resolution (i.e. 1024×1024), and the salient features of diverse faces were lost.

Our early experiments with StyleGAN[32] and StyleGAN2[33], however, showed promising results. Being style-based generators, they were able to map both macro (such as styles of different uniforms or backgrounds) and micro (such as facial attributes) features to a disentangled latent space. Also, by mixing the latent codes at both these levels, they were able to introduce diversity in synthesized images, which are key to visualize people of minority races and genders in different jobs. Therefore, these models hold significant promise to combat the bias problem in images in a new way. More importantly, both of these models can generate images at high resolution, which is a requirement for stock platforms. In our experiments, StyleGAN2 yielded better results than StyleGAN. Hence, we choose StyleGAN2 as our base network.

2.3.2 Introducing Explicit Control

Originally, StyleGAN2 was proposed to capture styles without supervision. But in our case, we would also like to exert some control over the generation process. Before we describe how we augment StyleGAN2 with multi-class conditioning, let us first briefly describe the basic structure of StyleGAN [32] the latent codes $\mathbf{z} \in \mathscr{Z}$ are first transformed to intermediate latent space $\mathbf{w} \in \mathscr{W}$ by a non-linear mapping network $f : \mathscr{Z} \rightarrow \mathscr{W}$. Then these $\mathbf{w}$ are transformed to "styles", $\mathbf{v} = (\mathbf{v_s}, \mathbf{v_b})$, which control the scale and bias in adaptive instance normalization operations (AdaIN) [23] after each convolutional layer of the generators of progressive GAN [30]. That is, $\text{AdaIN}(\mathbf{x}, \mathbf{v}) = \mathbf{v_s}[(\mathbf{x} - \mu(\mathbf{x}))/\sigma(\mathbf{x})] + \mathbf{v_b}$ where $\mathbf{x}$ is the feature map. Thus, the latent space $\mathscr{W}$ essentially controls styles within convolutional layers at each resolution through AdaIN. It is shown in [32] that these design choices for StyleGAN lead to a less entangled latent space in $\mathscr{W}$ compared to the input latent space in $\mathscr{Z}$. StyleGAN2 [33] further improves on this by redesigning its generator architecture and introducing a path length regularization into it to better learn the mapping from latent codes to images.

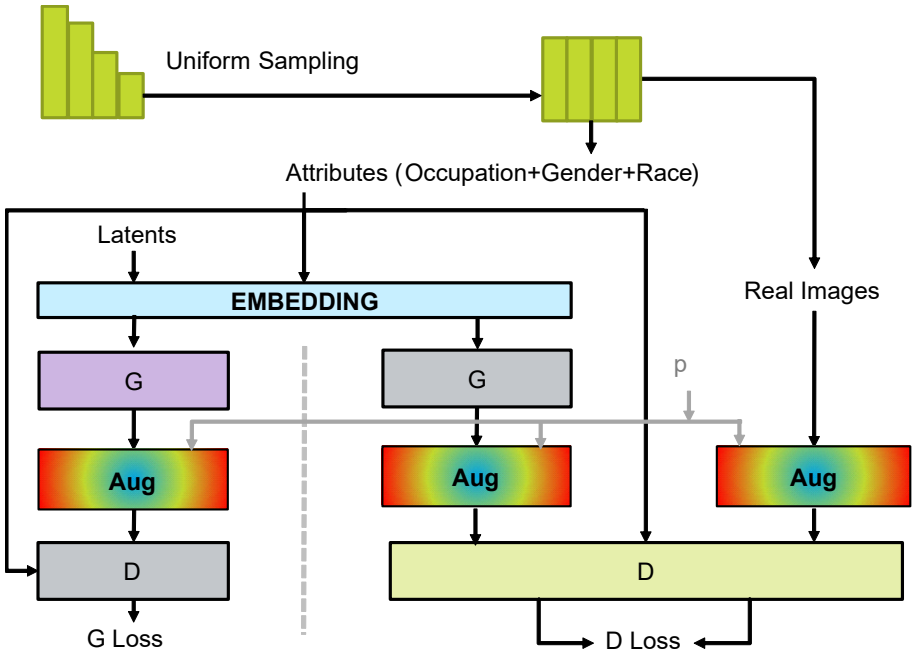


Figure 2.3. Model architecture and training framework for Uniform+. In addition to class-conditioning with regularization, we introduce new sampling techniques to handle the class imbalance (uniform sampling with augmentation).

Now, to explicitly control the generation process, we first one-hot encode each type of attribute (occupation, gender, and race) and concatenate them together into a single vector $\mathbf{y}$. That is: $\mathbf{y} = [\mathbf{y}_{\mathbf{occupation}}|\mathbf{y}_{\mathbf{gender}}|\mathbf{y}_{\mathbf{race}}]$. Then, we use a feedforward network to embed these features along with latent codes $\mathbf{z}$. The output of the embedding is then fed into the generator of StyleGAN2, G_θ. In our experiments, we have found that any other significant architectural changes to StyleGAN2's carefully designed generator lead to poor quality of images. For example, these following variations lead to mode collapse very early in the training procedures: (1) conditioning on mapped distribution $\mathcal{W}$ instead of random noise $\mathcal{Z}$, and (2) conditioning at each mapping layer from $\mathcal{Z}$ to $\mathcal{W}$. For the same reason, we do not apply any regularization to the generators like BigGAN [6] does. Rather, we make changes to the discriminator D_ϕ and apply the zero-centered gradient penalty from [47] to stabilize the high-quality conditional image generation process. Specifically, if $D_\phi(\mathbf{x}|\mathbf{y})$ is the discriminator score for an image $\mathbf{x}$ with condition $\mathbf{y}$, then the R regularizer is as follows: $R(\phi) = \frac{\gamma}{2}||\nabla D_\phi(\mathbf{x}|\mathbf{y})||^2$. Here, γ

is a hyperparameter to control the regularization process. To calculate the score from the discriminator, following their techniques, we have a separate real/fake discriminator for each class, and we predict separate logits for them. These discriminators share layers except the last layer, $f_\phi(\mathbf{x})$, which outputs a score for each class (thus, $f_\phi(\mathbf{x})$ has the same dimension as $\mathbf{y}$). Then, we perform an element-wise multiplication with our attribute vector to select the corresponding index for calculating the logit in the loss function. That is:

$$D_\phi(\mathbf{x}|\mathbf{y}) = \sum f_\phi(\mathbf{x}) \odot \mathbf{y} \tag{2.1}$$

We have also experimented using KL loss between attributes and predicted scores in the discriminator but found that it quickly leads to divergence. For training, we use the following non saturating loss [19] which is used in [33] for high-quality face generation from their Flickr-Faces-HQ dataset (FFHQ) [32]:

$$\mathscr{L}(\theta,\phi) = E_{p(\mathbf{z})}[f(D_\phi(G_\theta(\mathbf{z}|\mathbf{y})))] + E_{p(\mathscr{D}(\mathbf{x}))}[f(-D_\phi(\mathbf{x}|\mathbf{y}))] \tag{2.2}$$

where $\mathbf{x}$ is the input image, $\mathbf{y}$ is the attribute vector, and $f(t) = -\log(1+\exp(-t))$. This finalizes the design of our core architecture. We refer to this model as 'Vanilla.'

2.3.3 Combating Class Imbalance

Our vanilla model does not address the bias in the dataset that is needed to generate more examples with rare attributes. Unfortunately, similar to multi-class conditioning experiments, any major deviation from StyleGAN2's meticulously designed architecture to cope with bias leads to either poor results or training divergence. This motivates looking for alternative options and designing components that can be used to train any generative model to battle the bias problem. Below we describe each of them.

Weighted: Our first idea for improvement comes from cost-sensitive losses [17], where scores

for different classes are weighted to handle the imbalance of classes in the dataset in classification tasks. For weighting, we make the following changes in the output of the discriminator (Equation2.1) which is used in the loss function:

$$D_\phi(\mathbf{x}|\mathbf{y}) = \sum f_\phi(\mathbf{x}) \odot (\mathbf{m} \odot \mathbf{y}) \tag{2.3}$$

where $\mathbf{m}$ is the weight vector and has the same dimension as $\mathbf{y}$. If a class needs weighting, we set its corresponding index in $\mathbf{m}$ to an appropriate weight (described in 2.5.1). Otherwise, it is set to 1. We call this variation 'Weighted'.

Uniform: To explore another way to cope with bias, we note that the distribution of different attributes in the dataset is not uniform, but we can oversample from the dataset to create a uniform distribution during training. Another way to cope with rare categories in the data is to oversample from the dataset to create a uniform distribution during training. This can lead to better mapping of rare combinations of attributes. However, this may also lead to overfitting the discriminator, as the same images from rare classes appear more times and potentially destabilize the training. Our experimental results confirm this hypothesis. We observe that the FID score drops initially, but at a certain point it starts to increase continuously, and training begins to diverge. Nevertheless, we observed some improvement over Weighted. In the rest of the work, this variation is called 'Uniform.'

Uniform+: To stabilize Uniform, the key idea is to find a way of preventing overfitting due to the repetition of the same images from minor classes. Therefore, we hypothesize that augmentation can help if it can be applied appropriately within the discriminator. To overcome overfitting that arises from limited data, StyleGAN2-ADA [31] was recently proposed. They introduce Adaptive Discriminator Augmentation (ADA in short) which uses a wide range of augmentations with a probability $p < 1$ to prevent the discriminator from overfitting. They show that as long as the probability of a particular augmentation transformation is less than 1, the discriminator is still able to recover the original distribution. Given the effectiveness such augmentation to prevent

overfitting of the discriminator, we adapt it in our Uniform model to stabilize it. This leads to our final variation 'Uniform+' (shown in Figure 2.3). Our experiments show the effectiveness of this training procedure. We also adapt ADA in our vanilla architecture for comparison which we refer to simply as 'ADA.' It should be noted that the training procedure in Uniform+ is not specific to StyleGAN2 and can applied to other generative models as well.

2.4 Dataset

There is no existing dataset which we can use for our proposed task. Therefore, we have built a new dataset. In the following, we discuss how we have collected, preprocessed, and annotated the images in detail.

2.4.1 Collection

To obtain images for different occupations, we first construct our search query according to [34]. They conducted a study of which professions show the most racial and gender bias. From their list, we choose the following 14 professions: 'executive manager', 'administrative assistant', 'nurse', 'farmer', 'military person', 'security guard', 'truck driver', 'cleaner', 'carpenter', 'plumber', 'machine operator', 'technical support person', 'software engineer', 'writer.' We have collected around 10 thousand HQ raw images for these 14 occupations using Adobe stock API. We have chosen these 14 professions primarily because of their distinct styles or attires (we observed around 95% accuracy for top-3 prediction when we trained a classifier on them).

Table 2.1. Data statistics of Stock-Occupation-HQ. The breakdown shows the imbalance in race and gender across different professions.

Attributes	Exec. Mangr.	Admin. Asst.	Nurse	Farmer	Military	Security	Truck Driver	Cleaner	Carpenter	Plumber	Machine Op.	Tech. Support	Soft. Eng.	Writers
Male	447	68	297	659	164	374	488	211	379	582	338	127	199	195
Female	302	268	959	260	119	72	164	406	113	134	110	251	166	261
White	577	278	735	701	241	413	567	531	444	656	377	328	254	360
Black	96	12	263	59	26	25	40	50	12	37	25	22	42	41
Asian	38	38	163	106	10	6	35	29	17	12	32	17	30	38
Other	38	8	95	53	6	2	10	7	19	11	14	11	39	17
Total	749	336	1256	919	283	446	652	617	492	716	448	378	365	456

Figure 2.4. Real images from each profession after preprocessing

2.4.2 Preprocessing

A lot of images in this dataset are not ideal for training a generative model. First of all, many of the images do not have people in them. Even if an image contains humans, there can be multiple persons. Moreover, an image may contain complex backgrounds or complex foregrounds, which can make the task difficult for existing generative models to learn. Figure A.13 shows this. To overcome these challenges, we first use dlib's [37] face detector to detect faces, and then we use a custom padding scheme to crop the image around the face to include the upper body portion of each image. This overcomes the aforementioned challenges, i.e., keeping the problem simple while allowing critical information such as race, age, gender, and accessories/attire of different occupations intact. Still, we have noticed that a lot of images are not representative of the original occupation and contain generic photos. This required us to manually inspect the images and pick the best ones. After curating, the final dataset contains $8,113$ HQ (1024×1024) images in total. Figure 2.4 shows one example from each job, in the same order as the list above[3].

2.4.3 Annotating

To generate HQ images from attributes, we first need to label each image. For detecting gender and race automatically, we use a ResNet32-based [20] classifier that has been pre-trained

[3]Image attribution: okrasiuk, michaeljung, Günter Menzl, Andrey Popov, Kadmy, Piotr Marcinski, Kurhan, ronstik, michaeljung, and Al Troin on stock.adobe.com.

on the recently proposed dataset on fairness tasks: FairFace [29]. Its average accuracy is 95.7% for gender and 81.5% for race (Table 8 in [29]). Using it, we label each image for the following attributes, Sex: Male, Female, Race: White, Black, Asian, and Other Races. The overall statistics of our proposed dataset Stock-Occupation-HQ (SOHQ) is provided in Table 2.1. As can be seen from the table, the original distribution is highly imbalanced across different variables. This imbalance makes class conditioned image generation extremely difficult. Note that we do not introduce generic photos to increase diversity in each profession.

2.5 Experiments

2.5.1 Setup

Implementation details: We build our models in TensorFlow and we use the corresponding official codebase of StyleGAN2[4] and StyleGAN2-ADA[5] for base networks. As there are 14 professions, 2 genders, and 4 races, the attribute vector is 20 dimensional. For pretraining, we use two different datasets. First, we use StyleGAN2's pre-trained weights on the FFHQ dataset [32] in our Vanilla model, which we refer to as 'Vanilla-FFHQ.' However, images in our dataset are more challenging than FFHQ. This is because in addition to faces, our images contain various accessories, instruments, attires, and backgrounds related to the profession. Therefore, for pretraining purposes, we collected a large number of images (around 34 thousand) for 23 different professions and preprocessed them automatically (using our face detection and alignment pipeline). We call this dataset 'U-SOHQ,' for Uncurated Stock-Occupation HQ. We trained StyleGAN2 unconditionally on this dataset until convergence and use its pre-trained weights in all our models except for Vanilla-FFHQ. For Weighted, we set a weight of 2 for the 'Female' class, and 4 for the 'Black' and 'Asian' classes, based on their aggregated frequency. For ADA and the Uniform+ model, we set the probability of augmentation to 0.7. Finally, we set γ to 10 in the R regularizer (see Section 2.3.2) for all models.

[4]https://github.com/NVlabs/stylegan2
[5]https://github.com/NVlabs/stylegan2-ada

Metrics: For our first metric, we use the popular FID [22] score to quantify the quality of the generated images. FID measures the maximum distance between Gaussians fitted to the distributions of real and fake images. As the original distribution is biased, for a fair comparison, we sample attributes from the distribution of attributes in our dataset to generate images and then compute the FID with the real data. To measure how well the generated faces align with the given attributes, we measure the percentage of the time the given attributes match with the predicted ones. We call this metric 'Attribute Matching Score' (AMS). This is similar to classification error used in [11]. To predict the attributes from the generated images, we first generate 100 images each for all 112 combinations of race, gender, and occupation (11,200 images). Then, we detect race and gender using the classifier trained on FairFace [29]. For detecting profession we train a ResNet56 [20] on our dataset which achieves 80.28% top-1 accuracy (94.57% top-3). Using them, we compute the AMS for each attribute.

2.5.2 Quantitative Results

Table 2.2 shows the quantitative results for all models for all metrics. Under AMS, we show the matching scores for individual attributes ($G/R/O$) and the average of all three. On average, Uniform+ achieves the best results, although its FID is relatively high. "All 3" refers to the stricter criterion that all three attributes are correct simultaneously, and Uniform+ again has the best score. The FID score improves significantly between the Vanilla models when we use pre-trained weights from U-SOHQ instead of FFHQ. However, this results in lower attribute matching scores. We observe similar results from ADA (our conditional version of StyleGAN2-ADA [31]). While it achieves the lowest FID score, its combined AMS is the worst among all models.

We can explain this phenomenon as follows: let us assume that one model faces 'mode collapse' and thus outputs one image for each set of attributes. In this case, it is easy to generate an image that is faithful to the given attributes, so the AMS score will be high, but due to low variance in the images, the FID score will be high. On the other hand, imagine a model that

Table 2.2. Experimental results. Van-FFHQ: Vanilla-FFHQ, G: Gender, R: Race, O: Occupation. All models were pre-trained with U-SOHQ except Vanilla-FFHQ. The results show that Uniform+ achieves the best tradeoff between FID and AMS.

Model	**FID**↓	**AMS** (%)↑				
		G	**R**	**O**	**Avg.**	**All 3**
Van-FFHQ	21.11	**86.81**	38.66	63.66	63.04	23.71
Vanilla	14.89	86.00	34.72	60.14	60.29	20.34
ADA	**13.89**	80.78	34.76	67.79	61.11	19.99
Weighted	15.59	85.25	41.55	62.70	63.17	23.57
Uniform	22.75	85.30	43.77	**69.20**	66.09	27.21
Uniform+	17.34	83.33	**51.81**	63.48	**66.21**	**27.50**

produces diverse sets of background and styles of attires without being faithful to subtle facial attributes. In this case, it is possible to achieve a lower FID but the AMS will also decrease. Hence there is a tradeoff where a model has to achieve as low an FID as possible while keeping the AMS high.

Interestingly, Weighted comes close to achieving this goal. Its FID is lower while the attribute matching scores are higher. Uniform further improves on these matching scores. Unfortunately, Uniform has training divergence issues due to the repetition of the same images - that is, after reaching a minimum FID score, it starts increasing again as we continue training. The lowest FID score we were able to achieve for Uniform is 22.75, which is the worst of all the models. To rectify this, we introduced Uniform+, which uses augmentations from StyleGAN2-ADA [31]. We can see it achieves the highest combined AMS while keeping the FID score much lower than Uniform. Our training logs did not suggest any indication of divergence or mode collapse for Uniform+. Although its individual scores for gender and occupation are lower than Uniform, we will show in the following qualitative analysis that the performance gap is mainly due to similar images generated by Uniform.

Figure 2.5. Generated images from the four highest-scoring models show qualitative differences. The qualitative inspection reveals that Uniform+'s images achieve a better balance in all aspects.

Figure 2.6. Random examples from Uniform & Uniform+ for 'male white machine operator' query. This figure shows Uniform often generates similar-looking images (due to mode collapse). However, this is not the case for Uniform+.

2.5.3 Qualitative Results

For qualitative analysis, we use the best performing models under each metric, namely Vanilla-FFHQ, ADA, Uniform, and Uniform+. We evaluate their generalization performance by using the example queries from the introduction: 'female Black machine operator' and 'male Asian administrative assistant.' In our dataset, there is no image of the former, and just one image of the latter, so this is a strong challenge for the models. Figure 2.5 shows the results.

First, as can be seen in the Figure, all models struggle with these queries, as no model gets them all right. The first model, Vanilla-FFHQ, has relatively low variability, especially in

the female faces, as reflected in its FID score. It is able to generate correct 'female' and 'male' faces, reflecting its gender AMS. However, the generated faces (especially the males) are not racially correct, and the clothes do not appear to fit the intended occupation. ADA, on the other hand, shows a lot of variability in the generated images, but makes mistakes in all three attributes. Uniform is able to generate racially correct faces for both queries in most cases, but does not generalize well to the unseen query 'female Black machine operator' producing mostly male faces. On the other hand, Uniform+ generates images that are faithful to the given attributes, resulting in the highest combined AMS.

As we mentioned before, a model can perform better under AMS if it generates similar types of representative images for a query. We will now show that this is the case for Uniform but not for Uniform+. Figure 2.6 contrasts these two models. We observe that most images from Uniform have some artifacts in them, and similar types of images appear more than once (e.g., similar faces with yellow hats in similar orientations). This is clearly due to the repetition of images in its training set (note that we oversample in Uniform). This also shows early signs of mode collapse. Second, even though images are similar, the attributes are generally correct, so it has sacrificed diversity in the service of attribute accuracy. Unlike Uniform, Uniform+ is trained with more diverse images (due to augmentation). As a result, its images do not have artifacts or repetitions in them. This explains the performance gap we see between Uniform and Uniform+.

Human Evaluation. We have also performed a user study by hiring Amazon Mechanical Turk (AMT) workers to qualitatively evaluate the performance of the models. For this purpose, we choose ADA, Uniform, and Uniform+ and generate 100 examples for each of them using different queries. Based on our quantitative and qualitative results, we designed three different studies.

In the first study (called the Attribute Match Study), we ask evaluators to match the attributes of a generated image with the query that generated it, akin to the AMS metric. Since

Table 2.3. Human evaluation results (in percentage). For the last two studies, the percentage is calculated among the considered models.

Study Type		ADA	Uniform	Uniform+
Attribute Match Study↑	Gender	88.0	**89.0**	88.0
	Race	39.0	62.0	**69.0**
	Occupation	37.0	38.0	**44.0**
	All 3	11.0	14.0	**27.0**
Preference Study↑		20.2	23.0	**56.8**
Diversity Study↑		-	15.0	**85.0**

matching attributes does not capture the comparative quality of the images among the models (for example, see Figure 2.5), we performed a Preference Study. Here, we take one image from each of the three models, randomly shuffle them, and then we ask which image among the three is preferred by the evaluator for a query. Finally, in order to check for diversity of responses, we conducted a Diversity Study for Uniform and Uniform+ only. We presented a collage of 5 images from each of the models and we ask which one (after shuffling the order) generates more diverse images for a given query. Illustrative examples of each of these studies are provided in the supplementary material. We assign 5 unique Turkers for each task.

The results are presented in Table 2.3 and described below. For the Attribute Match Study, we report the percentage match. The majority vote among evaluators is matched to the attribute used in the query generating the image (this is similar to the automatic scores we calculated for AMS in Table 2.2). For the Preference and Diversity studies, we report the percentage of vote received by each of the models. The results from the human evaluation agree with our quantitative evaluation on Attribute Match. We see that Uniform+ gets the highest percentage of votes in most cases. As before, we can see Uniform has performed slightly better for matching one of the attributes (namely Gender). Previously, it performed better under AMS for matching Occupation. This indicates that Uniform performs better than Uniform+ in at least one aspect. While the numbers here are roughly consistent with the AMS scores, the Race scores are much higher and the Occupation scores much lower, revealing the inherent weakness of using automatic

Figure 2.7. Curated collection of generated images from Uniform+ (from left to right, top to bottom): 'exec. manager', 'admin. assistant', 'nurse', 'farmer', 'military', 'security', 'truck driver', 'cleaner', 'carpenter', 'plumber', 'machine op. ', 'tech. support', 'soft. eng. ', 'writer.' Zoom in for a better view.

metrics to evaluate images for this task. When images from all three models are presented side by side in the Preference Study, 56.8% of the time Uniform+'s images are preferred, which is more than twice as frequent as the other two models, again demonstrating the strength of Uniform+. Finally, when asked which model between Uniform and Uniform+ shows more diversity for a given query, Uniform+ received 85% of the total votes. This confirms our hypothesis that Uniform's occasional better performance is mostly due generating similar-looking images. Thus, based on our analysis, we find Uniform+'s performance strongest for our task. In Figure 2.7, we show curated examples of HQ generated images from Uniform+ where we pick one image across different combination of race and gender for each job.

2.6 Conclusion

In this work, we have proposed a new task of high-resolution image generation by controlling multiple attributes from imbalanced datasets to combat bias in image search. Our work makes several contributions to tackle new challenges for this task. First, we show how we can leverage existing state-of-the-art models for high-quality image generation and introduce explicit control over the generation process. Moreover, we show the challenges in training conditional models under a biased setting and propose new frameworks which can be applied

to any generative models by practitioners. We also produced a new, curated dataset as well as a large uncurated dataset for pretraining for the proposed task. Finally, we perform rigorous experiments that show the effectiveness of our proposed approach and reveal the weakness of the automatic metrics to gauge the quality of generated images for our task. We hope our design principles, as well as experimental studies, will benefit researchers to further improve on the models and propose new evaluation metrics for similar tasks.

Chapter 2, in full, is a reprint of the material as it appears in IEEE Winter Conference on Applications of Computer Vision (WACV), 2022. **Md Mehrab Tanjim**, Ritwik Sinha, Krishna Kumar Singh, Sridhar Mahadevan, David Arbour, Moumita Sinha, Garrison W. Cottrell, "Generating and Controlling Diversity in Image Search". The dissertation author was the primary investigator and author of this paper.

Chapter 3

Debiasing Image-to-Image Translation Models

3.1 Introduction

Generative Adversarial Networks (GANs) [19] have shown significant promise in synthesizing high-fidelity images [32, 33]. As a result, they have been adapted to achieve stunning results in many image-to-image translation (I2I) tasks, such as super-resolution [68, 39, 74], sketch-to-image [57, 10, 9], image inpainting [15, 75], etc. In these tasks, most current work focuses on the quality of generated results.

In this work, we study the capacity of existing image-to-image translation models to generate attributes that are in the minority in the training set. Figure 3.1 shows examples from the Pixel2Style2Pixel (pSp) model [55] network, which is one of the most popular and successful I2I models. The results for super-resolution and sketch-to-image tasks show an incredible visual quality of synthesized images, but also show an utter failure to generate minority visual attributes, such as eyeglasses (about 5% of the data) or baldness (about 2%), despite being clearly visible in the low-resolution or sketch input.

We have found that this problem is not limited to one particular architecture or dataset; whenever there is class imbalance in the training set, existing I2I translation models exhibit similar limitations. For example, we have trained the popular I2I translation model, pix2pix [26], on two synthetic datasets and found that the generated images are biased towards majority

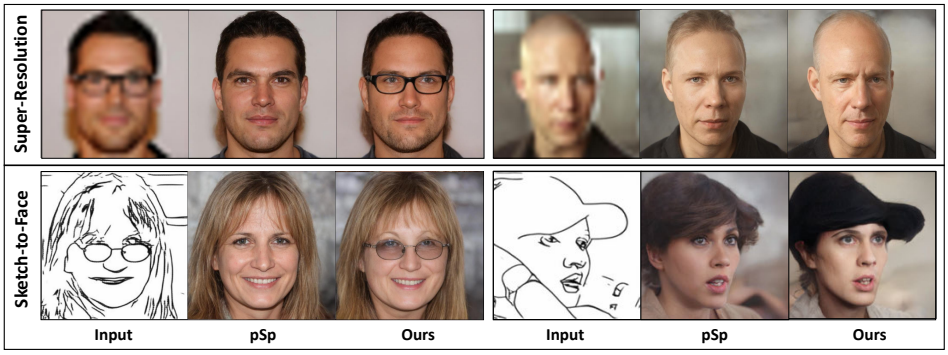


Figure 3.1. Examples of how Pixel2Style2Pixel (pSp) [55] is biased against minority attributes in the CelebA-HQ dataset [30]. We also show results from our debiasing framework (Ours).

attributes. This bias in I2I translation models can have a negative impact on various important downstream applications (e.g., image enhancement by super-resolution).

Here, we identify the need to debias I2I translation models, and propose a general framework to solve it. Normally used debiasing techniques, such as re-sampling [64, 7] and auxiliary classifier loss [50, 60], have not been explored for I2I. In this work, we showcase their success for I2I tasks. However, note that these methods only operate at the input level (re-sampling), or at final generation (pre-trained classifier). Without additional constraint in between, the latent features of biased classes can still become similar to the codes of the non-biased classes at the encoding level. This can prevent the decoder from learning a proper mapping between the latent codes and the output images from minority and majority classes during the generation process. To overcome this issue, we apply supervised contrastive learning during training to separate the encoded features of the minority from the majority, which helps the decoder to capture the features necessary to generate images with the correct attributes for both classes. We further conduct extensive experiments to show the effectiveness of our method on super-resolution and sketch-to-image I2I tasks. Figure 3.1 also shows how our method overcomes the bias problem for the pSp network. Note that our framework is agnostic to the particular encoder/decoder architecture.

Contributions:

1) We identify the bias problem in image-to-image translation and propose a new task of debiasing these models. 2) We propose a novel contrastive learning-based approach which outperforms the baselines both quantitatively and qualitatively. 3) Finally, we show that our model generalizes well, we apply it to multiple image-to-image translation tasks and datasets.

3.2 Related Work

Image-to-Image Translation. The goal of image-to-image translation models is to map images from a source domain (e.g., low resolution input) to images of a target domain (e.g., high resolution output), i.e., conditional image generation. The most notable work in this direction is pix2pix [26], where they show that conditional GANs can be used to solve a wide variety of image-to-image translation tasks. Motivated by their success, researchers have adopted specific conditional GAN architectures to solve specific image-to-image translation problems, for example, super-resolution [68, 39, 74], sketch-to-image [57, 10, 9], semantic label-to-image [52, 46, 81], unpaired translation [79, 45], or multi-modal image synthesis [12, 24]. However, most of these models are application-specific and may not generate high-resolution outputs.

For this problem, Pixel2Style2Pixel (pSp) [55] achieves promising results. Motivated by the capabilities of StyleGAN2 [33] to generate photo-realistic images, they train an encoder to project source images into the latent space of a pre-trained StyleGAN2 to solve image-to-image translation tasks. They show the effectiveness of their approach on tasks such as super-resolution, sketch-to-image, face frontalization, etc. Hence, we have chosen their model for conducting our experiments. Additionally, we have used popular pix2pix model [26] to test the generalizability of our approach.

Debiasing Frameworks. Bias and fairness have recently received a great deal of attention in the research community. Researchers have identified how the training datasets can suffer from various biases [66, 35, 65] and how it can lead to undesired behaviors in various image classification networks [61, 2, 21, 62, 14]. Some recent work explores using generative models

to create more balanced datasets for recognition tasks. For example, [4] creates synthetic images with latent space exploration so that bias in the classification network can be algorithmically measured. Similarly, to mitigate bias in classification networks, [3] adapted a variational autoencoder to learn the probability distribution of latent features. Based on the probability distribution, they re-sample those latent images which have lower probability to balance the dataset during training. There is limited work that aims to reduce bias for image generation itself; mostly, work has focused on the unconditional generation task to generate less biased distributions [58, 76]. All of these approaches focus on creating a balanced dataset, rectifying problems in the classification network, or doing unconditional generation.

In the case of debiasing image translation models, very few frameworks exist. [27] proposed debiasing I2IT models by using posterior sampling via gradient optimization, i.e., finding the optimal latent codes given the input. However, their application is limited in the case of reconstruction from noisy input, such as denoising or super-resolution. Furthermore, their debiasing method is not applicable to encoder-decoder architectures, which is the common architecture for I2IT models. Similarly, other works [69, 25] are either specific to model architectures, thus limiting the scope to apply their ideas to latest SOTA I2IT models, or need the generator to be retrained for learning a debiased representation. In this work, we propose a framework that can be applied to any I2IT tasks and models. Our proposed framework intercepts the encoding stage and it can even work for frozen generators (e.g., a frozen StyleGAN2 generator). Additionally, unlike all previous works, our framework can debias while maintaining high-quality. We describe our proposed framework in the following section.

3.3 Approach

In order to tackle the bias problem in the image-to-image translation models, we follow the common setting for the most of the debiasing work (e.g. [61, 2, 21, 62, 14]): we assume the bias is known or can be measured. In this section, we first discuss how we measure bias in the

Table 3.1. Bias analysis in the CelebA-HQ [30] dataset. The least three values in *Percentage* and *F1 Score on Generated* are shown in bold.

Attribute	Bald	Wearing Hat	Eyeglasses	Blond Hair	Bangs	Black Hair	Male	Heavy Makeup	High Cheekbones	Smiling
Percentage ↓	**2.37**	**3.57**	**4.89**	17.09	18.08	21.97	36.86	45.69	46.16	46.97
F1 Score on Real	0.8142	0.8908	0.9825	0.8483	0.8756	0.8186	0.9791	0.8906	0.8576	0.9333
F1 Score on Generated ↓	**0.7216**	**0.2500**	**0.0984**	0.8288	0.8393	0.7725	0.9653	0.8444	0.8235	0.9089

Celeb-HQ dataset for the image-to-image translation task. Next, we introduce our debiasing framework.

3.3.1 Measuring the Bias

We use two main criteria to detect bias. The first criteria is simple and straightforward: sort the attributes by the fraction of images in which they occur and select the attributes with lower fractions. Naturally, if images for a particular attribute are rare, then generation can become skewed against it. The second criteria is based on the ability to detect the attribute. This is important because we would like to reliably evaluate the performance of the model, and thus, we should select the attributes for which the classifiers show high accuracy.

To apply these criteria, we first train a ResNet152 classifier [20] on the same training set as Pixel2Style2Pixel [55] (pSp) and filter any attribute that shows a low F1 score (using a threshold of 0.8). For example, we observed an F1 classifier score higher than 0.95 for 'Eyeglasses' but a score of 0.0 for 'Blurry' which indicates 'Eyeglasses' clearly has better image feature representation than 'Blurry.' We therefore choose attributes that have high F1 scores and are rare in the dataset, providing us with rare attributes that are easily labeled automatically. To validate that being a minority in the training dataset can pose a bias problem for I2I task, we calculate the classifier F1 scores on generated images and assess the performance on minority classes. Table B.2 shows F1 scores on generated images along with the percentage of biased class and classifier F1 scores on the ground truth. Unsurprisingly, the under-performance is most pronounced in the attributes that are rare. There is a significant drop in the F1 scores between the real and generated images, when the attribute is in less than 5% of the images. We also see

this qualitatively in Figure 3.1. Additional examples are shown in the supplementary material. Here, we can see that none of the generated images from the pre-trained pSp network for either of the tasks faithfully reconstruct the attributes, although they are clearly visible in the input images. We also run the classifier on the input images to verify this (scores are included in the supplementary). This indicates the input images have information about the attributes and the pSp network should be able to reconstruct them. In the presence of bias, the model can also undesirably add some features. We can see from Figure 3.1, this is the case for 'Bald', where the model is hallucinating hair. These errors can have a detrimental impact on various downstream applications. Therefore, in this work, we address this problem and propose a general debiasing framework. For debiasing, we select the first 3 attributes from Table B.2, i.e.,'Bald', 'Wearing Hat', and 'Eyeglasses.' It is worth mentioning, although different tasks might require different ways of measuring bias, our debiasing framework is designed to be agnostic of how the bias is measured. In the following, we discuss our proposed debiasing framework.

3.3.2 Our Debiasing Framework

To mitigate bias in existing image-to-image translation models, the first insight comes from the exploration of images in the latent space. When the images from a certain group (e.g., images with eyeglasses or bald) are rare, their representation in the latent space can become similar to the majority group. As a result, the model becomes biased to frequent patterns. To remedy this problem, the key idea is to separate the latent codes for minority and majority groups, allowing the model to generate from their respective distributions independently.

A simple way to solve this during training is to over-sample the minority. Over-sampling the minority forces the network to see more instances of rare images and helps it encode the attributes. This is our first step towards debiasing is Step I in Figure 3.2. Although re-sampling is considered a general-trick for class imbalance problems, its effectiveness in image translation tasks has not been explored before. In some cases, it has been shown to not be effective [64, 63]. To improve on this, we propose using metric learning based losses [13, 71] to further separate

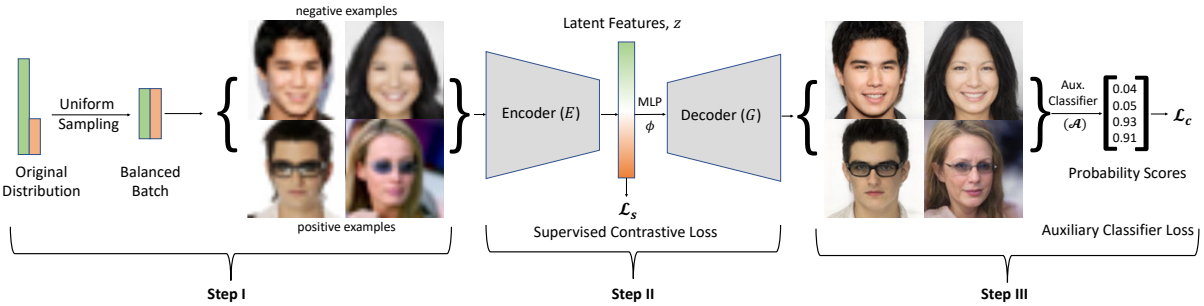


Figure 3.2. Our proposed debiasing framework. We first start by creating a balanced batch for a given attribute/class (Step I). Then, we apply supervised contrastive loss on the latent features (Step II). Finally, we apply an auxiliary classifier loss on the generated images (Step III).

the latent codes from different groups or classes. Here, we apply supervised contrastive loss [36]. This loss pulls together the representation of images from the same class (whether minority or majority) in the latent space and pushes them apart if from different classes. Mathematically,

$$\mathscr{L}_s = -\sum_{i \in \mathscr{I}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p)/\tau}{\sum_{x \in \mathscr{X}(i)} \exp(z_i \cdot z_x)/\tau} \tag{3.1}$$

Here, $i \in \mathscr{I} \equiv 1,\ldots,N$ (N is the batch size) is the index of an arbitrary sampled image I_i from the set of all images $\mathscr{I}$, $\mathscr{X}(i) \equiv \mathscr{I} / \{i\}$, $P(i) \equiv \{p \in \mathscr{X}(i) : y_p = y_i\}$ (y is the binary label or class of that image), $z_i = E(I_i)$ is the latent feature representation of the i^{th} image, I_i after it goes through the encoder E, and τ is a scalar temperature parameter. The positive pairs z_p in the supervised contrastive loss are obtained from the images that belong to same class and negative pairs are the images that belong to different classes. For example, images for 'Eyeglasses' will have positive pairs among themselves and images without 'Eyeglasses' will be the negative pairs in this case (shown as positive and negative examples in Figure 3.2). We should mention that we L_2 normalize the latent features to get the corresponding directions for applying supervised contrastive loss. However, the unit vectors or directions may not be suitable for generation. For this purpose, we pass the latent codes, z_i, through a multi-layer perceptron (MLP) layer, ϕ, after we have applied the $\mathscr{L}_s$ loss. The decoder or generator G then takes $\phi(z_i)$ as inputs for generating the target images. This is the second step in our framework (Figure 3.2 Step II).

Although we apply the contrastive loss in a supervised manner, and we are separating the latent codes of minorities and majorities, this may not always give the generator, G, enough incentive to focus on generating the particular attribute. So, to enforce this constraint further during the generation process, we use an auxiliary classifier $\mathscr{A}$ to predict the desired attribute on the generated images, $G(\phi(z_i))$, from the decoder and apply binary cross entropy loss:

$$\mathscr{L}_c = -\frac{1}{|\mathscr{I}|}\sum_{i\in\mathscr{I}} y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot(1-\log(\hat{y}_i)), \tag{3.2}$$

where $\hat{y}_i = \mathscr{A}(G(\phi(z_i)))$. This can further assist the supervised contrastive loss, $\mathscr{L}_s$, to separate the latent codes such that the desired attribute can be generated more easily. The final loss function is as follows:

$$\mathscr{L} = \mathscr{L}_o + \lambda_s * \mathscr{L}_s + \lambda_c * \mathscr{L}_c, \tag{3.3}$$

where $\mathscr{L}_o$ is the original loss function used to train the image-to-image translation model without our changes, $\mathscr{L}_s$ is the supervised contrastive loss, and $\mathscr{L}_c$ is the auxiliary binary cross entropy loss. The hyperparameters λ_s and λ_c balance the different losses. One thing to note from our debiasing steps is that our framework has no dependency on a specific encoder-decoder architecture. Thus, our approach generalizes to any image translation model.

3.4 Experiments

Dataset. We experiment on datasets where the bias occurs naturally. We also create two synthetically biased datasets.

1) CelebA-HQ. For experiments with human faces, we have selected the CelebA-HQ [30] dataset. As mentioned previously, in this dataset, the bias occurs naturally. The details on our train-validation-test split are described in the supplementary. *2) Bags and Shoes.* Our first synthetic dataset consists of images of bags from 'edge2bags' [26] and shoes from 'edge2shoes' [26]. We have selected a total of 5000 images, where 4950 images belong to 'Shoes' category, and the remaining 50 are from 'Bags' (99:1 bias ratio). We call this dataset 'Bags and Shoes.'

We separately keep 200 images in total for both validation and test set. *3) Cats and Dogs.* For this dataset, we select animal faces from AFHQ [12]. Specifically, we have selected faces of Cats and Dogs. The training, validation, and testing split follow the same strategy as 'Bags and Shoes.' In this dataset, the majority class is 'Cats.'

Tasks and Models. For experiments with faces, we select two popular image-to-image translation tasks, namely, super-resolution and sketch-to-face. For performing these translation tasks on human faces, we have selected one of the recently proposed image-to-image translation models, Pixel2Style2Pixel (pSp) [55]. As debiasing the image-to-image translation task is new, there are no existing baselines. Hence, for comparison, we created our own baseline and variants of our model: **1) Vanilla.** Original pSp network without any changes. We train the network from scratch on our dataset for each of the attributes. **2) Sampling Baseline.** This is a trivial baseline where images from the minority class are resampled to create a balanced batch (our Step I). We also apply data augmentations (e.g. shifting, shearing, scaling, horizontal flipping, etc.) to all images when re-sampling. **3) Ours (I+II)** In this model, we take the first two steps from our pipeline, that is re-sampling and applying supervised contrastive loss (Equation 3.1), $\mathscr{L}_s$, during the encoding-decoding phase. **4) Ours (I+III)** Here, we only consider re-sampling and applying auxiliary loss (Equation 3.2), $\mathscr{L}_c$. **5) Ours (I+II+III)** All three components in our debiasing framework.

For all of our experiments with pSp, we use the same frozen StyleGAN2 as the decoder, pre-trained on FFHQ (which has good coverage for accessories like eyeglasses, hats, etc.). Using this network, we were able to generate rare features in Celeb-HQ such as eyeglasses, hats, etc. (Figure 3.1, 3.3). This shows that the latent codes are already available in the pre-trained StyleGAN2 generator network. Any problems, therefore, lie in pSp's encoder, which becomes biased during training with a biased dataset (e.g. Celeb-HQ).

Our debiasing framework is not only limited to the pSp network and human faces. We can apply it to different I2I translation architectures, and images from domains other than human faces. To show this, we have chosen another popular image-to-image translation network, pix2pix

[26], and perform edge-to-image task on our synthetic datasets, namely 'Bags and Shoes', and 'Cats and Dogs.' Similar to pSp, we create different variants of our model for pix2pix. More details of our data augmentations and training procedures for both models can be found in the supplementary.

Evaluation. To measure how well the models faithfully generate the attribute (or absence of it), we report the classifier prediction scores on the generated images. We convert all scores to the same scale (between 0 and 100). A model is better if it obtains high scores on the minority group while maintaining the majority group performance. For fairness, we keep the classifiers for evaluation different from our models (more details in the supplementary). For the super-resolution task with pSp, we also report Learned Perceptual Image Patch Similarity (LPIPS) [77] and MSE in order to evaluate whether our generated images overall match the target images. For experiments with pix2pix, we perform edge-to-image synthesis; given the loss of information in this task, we do not consider there to be a 'ground truth' image, so LPIPS/MSE do not apply. In this case, we report Fréchet Inception Distance (FID) [22] to measure if the generated images match the actual distribution of their respective classes.

3.4.1 Quantitative and Qualitative Results

Table 3.2 shows the results. Our model and its variations always outperform the Vanilla and Sampling baselines for the minority groups. In terms of majority performance, Vanilla's performance is often better, as one would expect since it is biased to the majority class. The table also reveals the individual contribution of all three components of our model. For example, applying only supervised contrastive loss on top of re-sampling (**I+II**) helps in almost all cases. Applying the auxiliary classifier loss (**I+III**) helps even more in 11 out of 18 cases. But when supervised contrastive loss is applied with the auxiliary classifier (**I+II+III**), it improves the minority class scores for almost all cases, with a negligible difference in the remaining case. For example, it leads to about 30% and 57% improvement in super-resolution and sketch-to-face, respectively, for the 'Wearing Hat' minority class compared to the sampling baseline. In this new

Table 3.2. Comparison of classifier prediction scores on all groups among the models across different tasks and attributes.

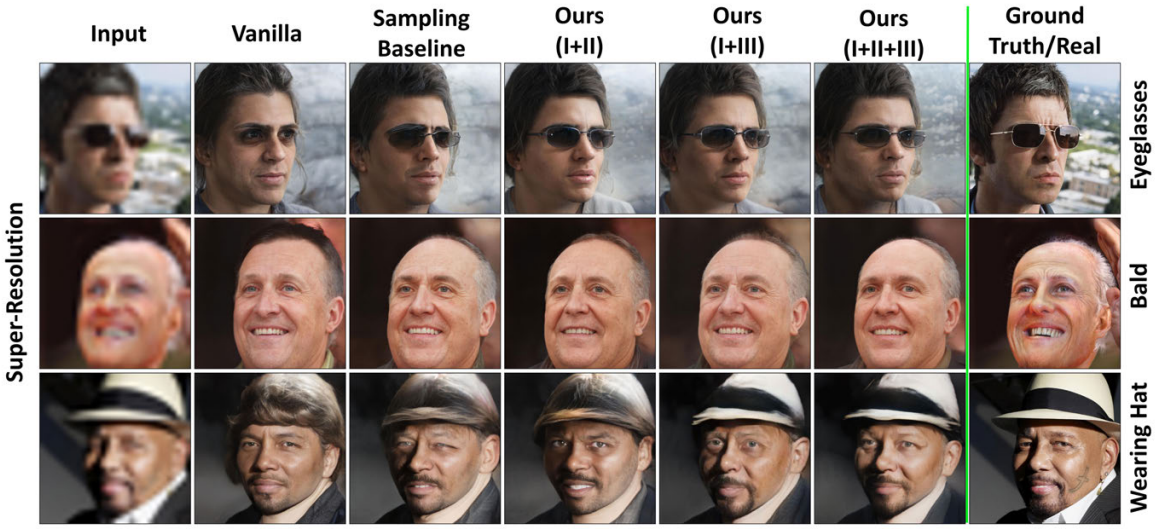
Task	Attribute	Group	**Vanilla**	**Sampling Baseline**	**Ours (I+II)**	**Ours (I+III)**	**Ours (I+II+III)**
Super-Resolution	Eyeglasses	Minority	15.27	89.95	91.55	92.35	**92.85**
		Majority	98.52	98.21	98.6	97.8	**98.7**
		Both	56.89	94.08	95.08	95.08	**95.77**
	Bald	Minority	63.91	88.89	90.41	90.12	**91.60**
		Majority	**98.18**	97.53	98.01	98.01	98.01
		Both	81.04	93.21	94.21	94.06	**94.81**
	Wearing Hat	Minority	23.19	60.71	61.95	78.31	**80.84**
		Majority	**98.4**	97.62	97.93	97.95	98.21
		Both	60.8	79.17	79.94	88.13	**89.52**
Sketch-to-Face	Eyeglasses	Minority	30.32	92.73	93.30	**94.10**	94.06
		Majority	98.65	96.05	97.99	98.7	**98.84**
		Both	64.48	94.39	95.64	96.39	**96.45**
	Bald	Minority	46.88	85.26	81.03	86.80	**89.12**
		Majority	**98.3**	95.47	97.09	96.46	95.53
		Both	72.59.	90.36	89.06	91.63	**92.32**
	Wearing Hat	Minority	15.58	32.38	50.49	78.61	**79.50**
		Majority	**98.37**	98.07	97.69	97.01	97.19
		Both	56.98	65.22	74.09	87.81	**88.34**

Table 3.3. Quantitative results for image reconstruction in the super-resolution task. Our approach does not compromise image quality.

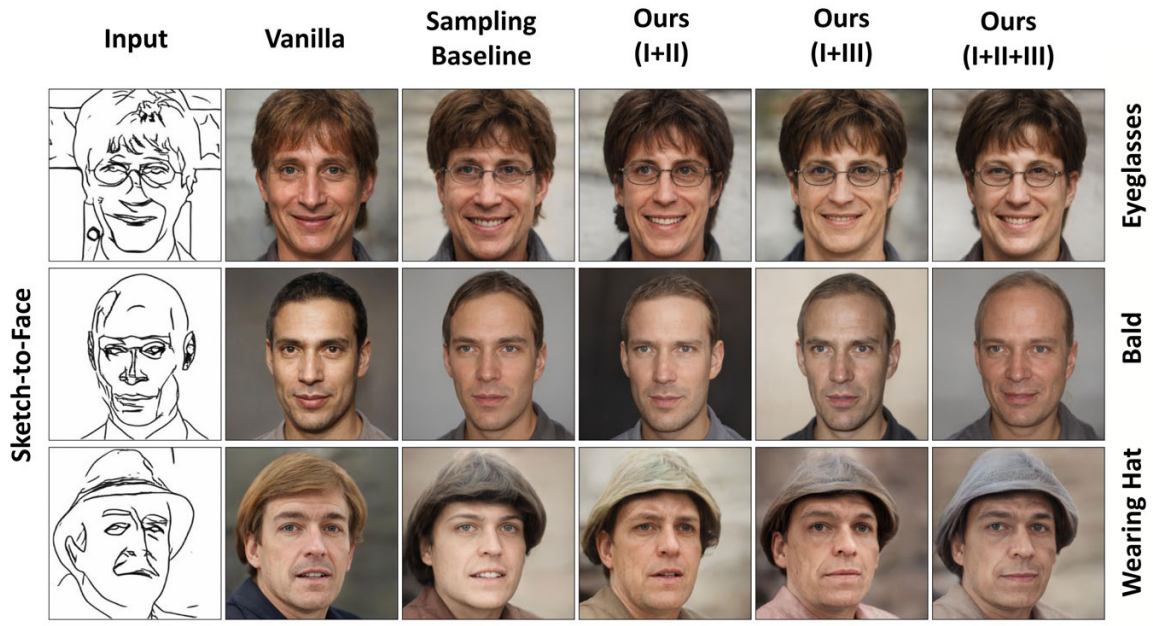
Metric	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)
LPIPS↓	0.25 ± 0.06	0.24 ± 0.06	0.25 ± 0.06	0.24 ± 0.06	0.25 ± 0.06
MSE↓	0.06 ± 0.03	0.05 ± 0.03	0.06 ± 0.03	0.05 ± 0.03	0.06 ± 0.03

task, general tricks like re-sampling and auxiliary classifier help a lot, which is not always the case in many debiasing tasks [63, 70, 64, 72], but adding the contrastive loss generally improves on these.

For tasks like super-resolution, it is also important to match the quality of generated images with the ground truth. Therefore, we report LPIPS [77] and MSE in Table 3.3. We can see our framework performs debiasing without compromising the image quality. For further qualitative comparison, we show the generated images from each of the models for each attribute and task in Figure 3.3. We also show the ground truth images for super-resolution in the right most column of 3.3 (a). We can see, in all cases, the Vanilla model does not produce the desired outputs. Compared to other alternatives, we can see our methods produce much better results.



(a) Super-Resolution



(b) Sketch-to-Face

Figure 3.3. Results of our debiasing framework compared to the Vanilla and Sampling Baseline model. Here, we show one example for each of the considered tasks across all attributes. Our generated results better capture the attributes compared to baselines.

Table 3.4. FID scores show the effectiveness of our approach in a different image-to-image translation architecture, using images from different domains.

Group	Bags and Shoes					Cats and Dogs				
	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)	Vanilla	Sampling Baseline	Ours (I+II)	Ours (I+III)	Ours (I+II+III)
Minority	202.30	135.22	140.26	130.73	**122.77**	241.98	189.35	181.29	183.04	**177.22**
Majority	233.92	159.33	151.63	152.66	**116.89**	**64.32**	70.40	76.22	65.82	68.42
Both	195.29	130.16	128.51	124.66	**105.18**	136.86	116.92	116.24	111.99	**110.7**

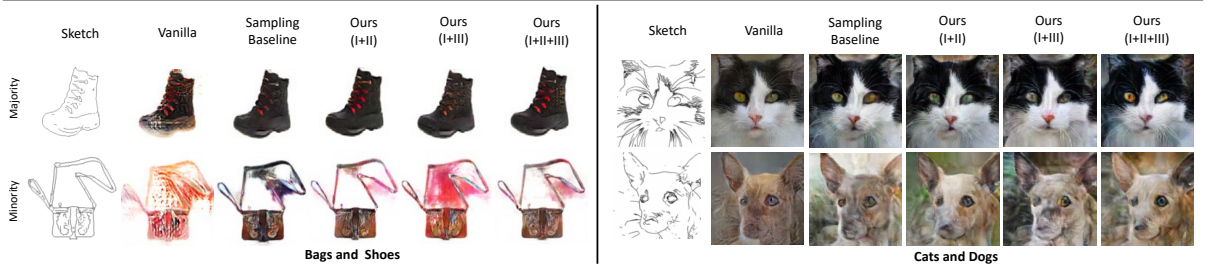


Figure 3.4. Our debiasing framework is not only limited to a particular model. Here, we show how our idea can be applied to pix2pix [26] to improve the quality of synthetic images in the presence of bias.

For example, 'Wearing Hat' appears to be the hardest attribute to reconstruct, for both tasks. Even for this attribute, we can see our contrastive model (I+II+III) is producing hat-like shapes and textures. Additionally, we can see the quality of the images from our model is similar or better than Vanilla in all cases, which suggests that our framework is an important tool for image-to-image translation models in the presence of bias.

3.4.2 Generalization to a Different Architecture

Here we discuss our results when we apply our framework to pix2pix [26] on datasets other than human faces. Figure 3.4 shows the qualitative results among the models. For both datasets, a common pattern is that the quality of the majority does not change by much. The quality of the minority, however, varies a great deal. For 'Bags and Shoes', we can see both Vanilla and the Sampling Baseline model try to fill the gap between the body of the bag and strap. This is because most of the images from this dataset are shoes, and the contour of shoes are always filled. Therefore, to resemble the majority ground truth images, the generative model tries to fill the gap. The contrastive learning based approaches, especially (I+II+III), do not fill up the space between strap and bag as much, and show minimal changes compared to the

other models. Similarly, for 'Cats and Dogs', the majority class is 'Cats,' and a frequent pattern is images having cat-like fur. As a result, both Vanilla and Sampling Baseline's outputs have cat-like fur in the minority group's ('Dog') images. However, as can be seen from this figure, our (I+II+III) model's coloring is more authentic for the 'Dogs' class. We have also quantitatively evaluated the performance of all the models by calculating FID scores between the generated images and ground truth. Table 3.4 shows the results. As we can see, our model achieves the lowest scores, especially for the minorities, indicating better quality of images for the selected task. For example, adding contrastive loss to re-sampling with auxiliary classifier leads to 18% reduction of overall FID for 'Bags and Shoes' dataset.

Overall, our contrastive-learning based framework leads to consistent improvement for both pSp and pix2pix (Table 3.2 and 3.4, Figure 3.3 and 3.4).

3.5 Discussion and Limitations

Our debiasing framework achieves better performances compared to other alternatives across different categories, different architectures, and different domains. However, so far, we have made an assumption that the bias is known, which is the most common assumption in many debiasing work [61, 2, 21, 62, 14]. This is because as long as there are attributes in the dataset, we will always be able to know if the dataset has class imbalances and whether that might lead to bias in the model. There can be various ways to measure the bias, and in Section 3.3.1, we explored one way of measuring it.

In this work, there is another inherent assumption that the dataset is biased towards only a single attribute/class. In reality, bias can appear in multiple attributes/classes simultaneously. For example, Figure 3.5 shows minority attributes, namely 'Bald' and 'Eyeglasses', appear in the same image. The figure also shows how focusing on only one attribute might not necessarily debias it for the other one (i.e. debiasing for 'Bald' does not debias for 'Eyeglasses', and vice versa). One simple, straightforward way to tackle this problem will be to merge multiple labels into single classes and apply our debiasing framework. However, doing so might not be scalable

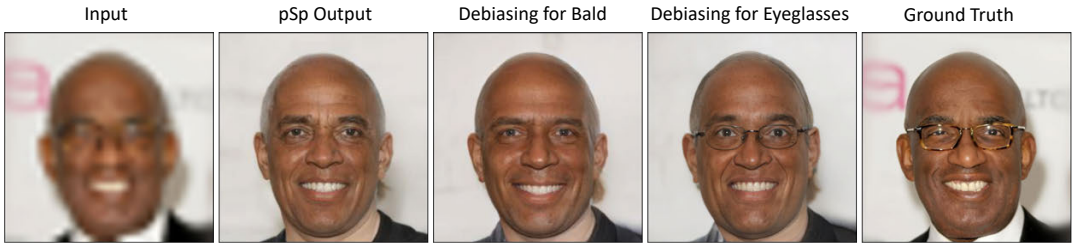


Figure 3.5. An example case where dual bias can appear. In this example, the ground truth image has both 'Bald' and 'Eyeglasses' attribute, and debiasing for only one attribute does not necessarily debias for the other one.

if the labels are large in number. We would like to explore this direction in future.

3.6 Conclusion

In this work, we propose the new task of debiasing image-to-image translation models. Using Pixel2Style2Pixel and pix2pix, we have demonstrated that minority attributes are poorly reconstructed whenever there is an imbalance in the dataset. To solve this problem, we have proposed a novel contrastive-learning based approach to separate the latent codes of minority classes from the majority classes. From the experimental results from both pSp and pix2pix, we have shown that this contrastive learning approach, when coupled with general tricks like re-sampling and auxiliary classifiers, leads to consistent improvements across all the tasks. Our framework does not depend on any particular translation model or dataset, making our solution model and data agnostic.

Chapter 3, in full, is a reprint of the material as it appears in British Machine Vision Conference (BMVC), 2022. **Md Mehrab Tanjim**, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell, "Debiasing Image-to-Image Translation Models". The dissertation author was the primary investigator and author of this paper.

Chapter 4

Discovering and Mitigating Biases in CLIP-based Text-to-Image Generation

4.1 Introduction

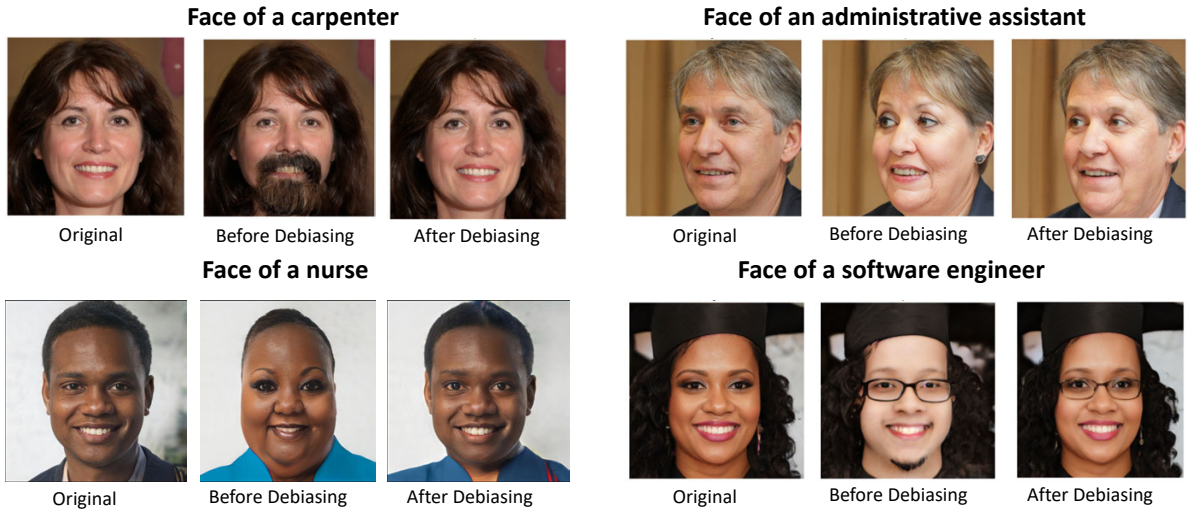


Figure 4.1. Biases in the CLIP model [54] can bias CLIP-based text-to-image generation. Here, examples are shown from a CLIP-based generator, StyleCLIP [53]. We also show our debiasing results using our proposed techniques.

CLIP (Contrastive Language-Image Pre-Training) [54] is a neural network trained on a large set of pairs of images and texts. It has excellent zero-shot capabilities, e.g., it matches the performance of the original ResNet50 [20] on ImageNet without explicitly getting trained on the original labels. Due to its rich learned features between text and image modalities, recently it has

been showing significant promise in text-to-image synthesis as well. For example, StyleCLIP [53] leverages the text-image associations of CLIP models and the generative power of StyleGAN [32] to develop a free form text-based interface for image editing.

Despite being trained on a large dataset, it has been shown that CLIP models suffer from various biases [1]. However, these studies have focused on the biases in classification tasks. To the best of our knowledge, no studies have been done on how biases in CLIP impact generative models. In this work, using StyleCLIP as a case-study, we reveal biases in CLIP and show how these biases negatively impact the generation process. Figure 3.1 shows some illustrative examples. Here we show the original images as well as the manipulation by StyleCLIP for the given text prompts. The images clearly show the generated images suffer from gender and racial biases, which have negative societal impact.

In this work, we also propose several ways to mitigate the biases. We have identified that the text CLIP embedding has learned correlations between different occupations and gender or race. For example, 'a nurse' has high similarity with 'a female'. To debias the text embedding, we remove the gender and race component from the text queries. We have found such techniques can debias simple cases such as face editing. However, it does not sufficiently debias complex cases (e.g., change of occupation-related clothing). For such cases, we introduce a gradient-based optimization which provides resistance against biases based on identity preserving losses (by calculating LPIPS [78] score or the cosine similarity between two faces using ArcFace [16])). Our optimization still provides enough incentives to make necessary changes for the given text prompt (based on the CLIP loss). Our debiasing framework does not require retraining CLIP or the generative model (e.g., StyleGAN in case of StyleCLIP). Also, our techniques do not depend on particular query words (e.g., occupations) and can be generalized to debias other cases of biases. Images shown as 'After Debiasing' in Figure 3.1 demonstrate the debiasing capabilities of our framework.

Contributions:

1) We discover the biases in the CLIP model and, using StyleCLIP, we show their negative impact on text-to-image generation. 2) We propose several techniques to debias without retraining CLIP or the generator.

4.2 Related Work

Bias and fairness have recently received a great deal of attention in the research community. A common way to tackle the bias problem is to train on a large balanced dataset. So, one can imagine the bias will not be prevalent in large, pretrained models like CLIP [54]. However, researchers have audited this model for various classification tasks and discovered biases [1]. These studies focused on the bias problem with respect to classification. However, CLIP models are being widely used to train various generative models as well. Any bias in CLIP models can therefore negatively impact the generation process. For this reason, here we focus on discovering biases in CLIP models for generative tasks. For the generative model, we have chosen one of the most popular CLIP-based generators, StyleCLIP [53]. To the best of our knowledge, ours is the first work to discover any biases in CLIP-based models for generative tasks.

4.3 Approach

Discovering Biases. To discover biases, similar to other recent work [64], we collect stock images for different occupation-related queries. We chose the following professions: 'Plumber', 'Nurse', 'Administrative Assistant', 'Farmer', 'Security Guard', 'Executive Manager', 'Military Person', 'Maids & Housekeepers'. For each of them, we add Male/Female/White/Black in the beginning to construct search queries. In this way, we have two sets: one for gender (total 332 images) and one for race (total 379). Then we rank these separate sets of images according the CLIP score using these different professions as queries. Ideally, images belonging to a profession should be ranked higher than others if that profession is used as a query, regardless of race and

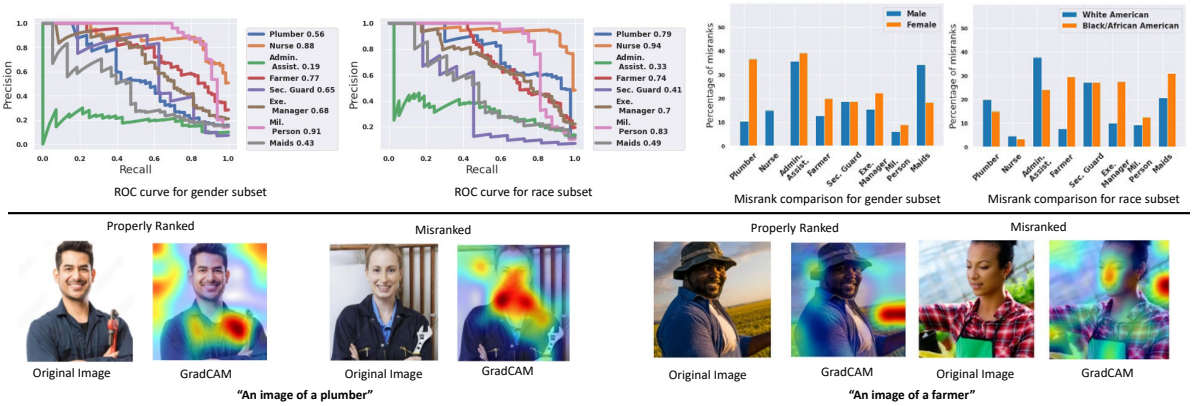


Figure 4.2. (Top) We show the ROC curves and percentage error comparison in ranking stock images using CLIP scores for different occupation-related queries. (Bottom) GradCAM shows where the CLIP model focuses for a given text prompt.

gender. However, that was not always the case. In Figure 4.2, we show the ROC curve for both the gender and race image set. We can see the performance is quite low for most of the queries, despite the fact that the CLIP model usually has excellent zero-shot performance. We also show the percentage of times the model misranks for a specific gender/race. These figures show, when it misranks the images, certain genders and races are misranked more than others. For example, we can see for plumbers, female plumbers are often misranked. Figure 4.2 also shows some examples with GradCAM visualization [59]. GradCAM shows, for the male plumber image, the CLIP model correctly focuses on instrument but for the female plumber, it focuses on her face, resulting in a misrank. Similarly, for farmer, in both cases the model focuses on the green background. However, it gives an additional focus on female faces, causing a misrank. This has a negative impact on text-to-image generation (Figure 3.1).

Our Debiasing Framework. We have identified that most of the occupation queries have high similarities to a particular gender or race in the text embedding space in CLIP. So, taking out the gender/race component from the text can potentially debias the generation. To do so, we follow similar approach as [5]: we take commonly used male-female words and perform PCA to find gender directions. Then for any text embedding, we project it onto the gender direction and take the orthogonal direction to it to remove the gender component. This technique proves to be

quite effective. For example, the images for 'Face of a carpenter' and 'Face of an administrative assistant' from Figure 3.1 are debiased in this fashion.

Unfortunately the previous method does not work well for all cases, especially where there are complex changes required for the given text prompts. For example, in Figure 4.3, we show a generated image of a nurse using the StyleGAN-based occupation generator from [64] (shown as 'Original' in leftmost). This is a complex image as it has the uniform and equipment of a nurse, unlike images of only faces. For this image, we load the generator from [64] into StyleCLIP and prompt it with 'A plumber' text. The output from StyleCLIP is shown next to the original image. We can see the impressive capability of StyleCLIP to modify the attire and background of the original image to portray a plumber. Unfortunately, due to bias in CLIP, the generated image shows an image of a male plumber. In this complex case, we can also see our previous text-based debiasing did not produce a satisfactory result (shown as 'Text-based Debiasing'). We can see, except for the face, everything else remained almost the same as the original image. To improve on the text-based debiasing in these cases, we introduce a gradient-based latent code optimization. Mathematically, if s is the original latent code, G is the generator, then we find a new latent code by $G(s \pm \alpha)$. Here, the α is initialized with zeros and optimized via back-propagation from the following loss: $\mathscr{L} = \mathscr{L}_{CLIP} + \beta_1 * \mathscr{L}_{Gender} + \beta_2 * \mathscr{L}_{ID} + \beta_3 * \mathscr{L}_{LPIPS}$. Here we use the CLIP model to determine the similarity of the generated image with given text prompt, and subtract it from 1 to get the CLIP loss, $\mathscr{L}_{CLIP}$. For $\mathscr{L}_{Gender}$, we use a gender classifier to determine the difference between gender prediction scores between the original image and the generated image. We calculate the ID loss, $\mathscr{L}_{ID}$, by calculating the distance between the original face and generate face using ArcFace [16]. LPIPS loss, $\mathscr{L}_{LPIPS}$, is calculated using [78]. The β values are hyperparameters to control the effect of different losses. In Figure 3.1, for 'Face of a nurse' and 'Face of software engineer', we have used our gradient-based optimization to debias as text-based debiasing did not produce satisfactory results for these prompts. For these examples, using CLIP and ID loss was sufficient. For the image in Figure 4.3, we show all combinations of losses and using 3 out of the 4 losses appears to work best. From both Figure 3.1

Figure 4.3. Our gradient-based debiasing framework with different combinations of identify preserving losses. Here, the text prompt for StyleCLIP is: 'A plumber'.

(bottom row images) and 4.3, we can see our proposed gradient-based latent code optimization successfully debiases the difficult examples. Note that our proposed framework does not require retraining the CLIP or the StyleGAN generator model.

4.4 Conclusion

The CLIP model is widely being used for various tasks. It is therefore important to address if any bias in CLIP negatively impacts the given task. In this work, we have discovered such biases in text-to-image generation. We have also demonstrated methods that mitigate these biases without retraining CLIP or the generative model.

Chapter 4, in full, is a reprint of the material as it appears in Responsible Computer Vision at European Conference on Computer Vision (RCV@ECCV), 2022. **Md Mehrab Tanjim**, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell, "Discovering and Mitigating Biases for CLIP-based Text-to-Image Generation". The dissertation author was the primary investigator and author of this paper.

Chapter 5

Conclusion

Generative models have shown a lot of promise, especially in the domain of images. The quality of generated examples from GANs exceeds any of the previous image generative models, and with the availability of GPU resources and a large amount of data, GANs are able to achieve unprecedented image quality. However, little attention has been paid to how biases in the training data or models can affect image results. Debiasing image generative models is therefore a critical area of research as it helps to ensure that these models are more equitable and unbiased in their outputs.

This dissertation focused on exploring various approaches to debiasing image generative models and evaluating their effectiveness. Initially, our research delved into the systemic biases that have led to certain professions being more prevalent among specific genders and races. These biases are also apparent in the search results of stock image repositories and search engines, presenting a challenge for content providers. Given the limited choices of existing content for particular combinations of profession, race, and gender, it is crucial to provide content users with the ability to depict a broad range of professions with diverse racial and gender characteristics. Our research aims to address these issues and contribute to a more impartial and equitable future. As a result, we introduced a new task of high-fidelity image generation from imbalanced datasets that takes multiple attributes into account.

In our second study, we explored the issue of bias in image-to-image translation models

when they are trained on imbalanced datasets. Our research demonstrated that using contrastive learning, in addition to other commonly used techniques like sampling and auxiliary classifier loss, can help to alleviate this problem. While our current findings pertain to a single attribute at a time, it is possible to extend this approach to address biases for multiple attributes simultaneously by creating distinct classes for each attribute.

In our last study, we focused on examining biases in large pretrained models like CLIP, and demonstrated how such biases can adversely affect text-to-image generative models like StyleCLIP. To address this issue, we put forth identity-preserving losses that can effectively mitigate the problem without the need for retraining the pretrained model.

This dissertation primarily focused on conditional generation, specifically attribute, image, and text-to-image generation, and proposed a debiasing framework. However, beyond these conditional settings, it is also possible to extend the proposed debiasing framework to unlabeled data. One approach to achieving this would involve learning latent variables during generation using mutual information maximization [51] and resampling based on the learned probabilities, thereby increasing the likelihood of generating rare images with distinct latent features [3]. Our debiasing framework can also be extended to generative models beyond GANs, for example, Stable Diffusion Models [56]. However, this remains a topic for future research.

In conclusion, this dissertation presents numerous promising avenues for debiasing generative networks. Our proposed techniques for debiasing image generative models offer the potential to ensure that their downstream applications across various products, such as Image Stock platforms or photo-editing software, do not produce any unwanted outcomes.

Appendix A

Generating and Controlling Diversity in Image Search- Supplementary Material

A.1 Human Studies

We show one illustrative example from each of our human study types in the next page. Specifically, we show an example task from our Attribute Match Study in Figure A.1, an example from our Preference Study in Figure A.2, and an example from our Diversity Study in Figure A.3.

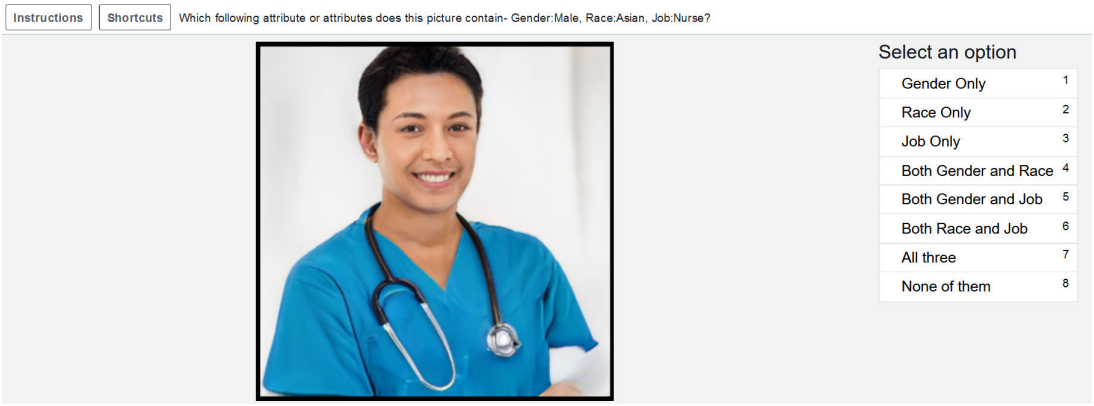


Figure A.1. An example task from our Attribute Match Study. Here, the image is generated from Uniform+.

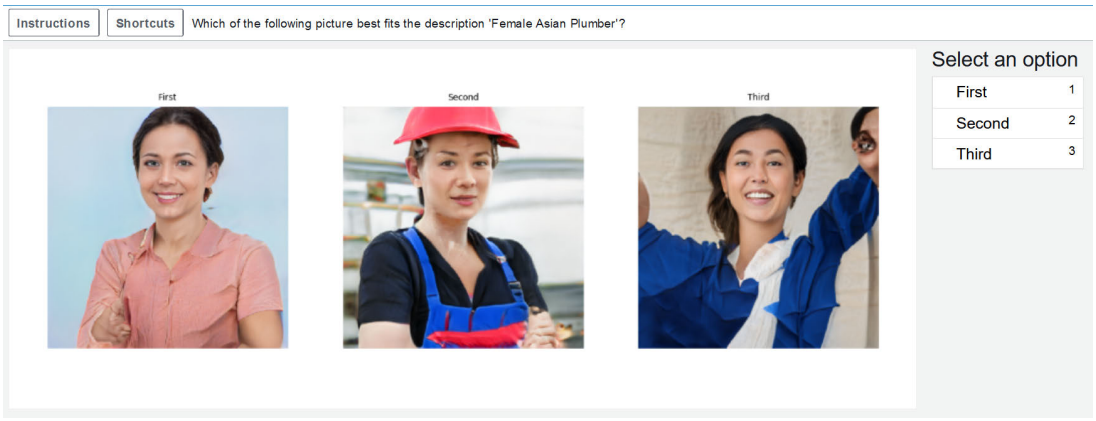


Figure A.2. An example task from our Preference Study. Here, the models are as follows- First:Uniform, Second:Uniform+, Third:ADA.

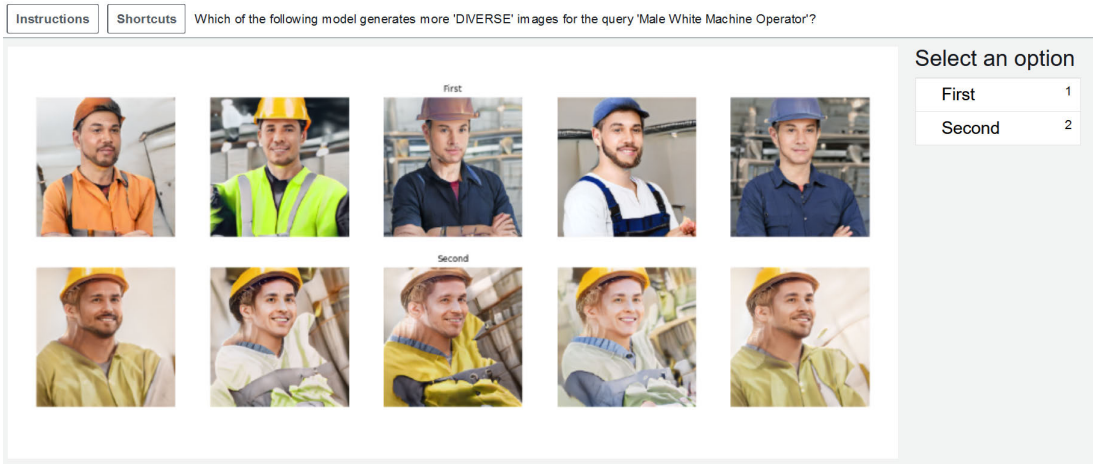


Figure A.3. An example task from our Diversity Study. Here, the models are as follows- First:Uniform+, Second:Uniform.

A.2 Additional Generated Images

We also show one generated image from our strongest model Uniform+ for each combination of races and genders for each of the 14 professions in the following pages (Figure A.4 - A.17). These images show the potential of Uniform+ to combat the bias in image search. For generating pictures, we used the truncation trick, where the center of mass in the latent codes $\mathbf{w}$ is first calculated, and then samples are chosen randomly within a deviation ψ (called the truncation value) from this center. Mathematically: $\mathbf{w}' = \bar{\mathbf{w}} + \psi(\mathbf{w} - \bar{\mathbf{w}})$, where $\bar{\mathbf{w}}$ is the center and the $\mathbf{w}$'s are the latent codes. We observed good results when we set $\psi = 0.7$.

Figure A.4. Executive Manager

Figure A.5. Administrative Assistant

Figure A.6. Nurse

Figure A.7. Farmer

Figure A.8. Military Person

Figure A.9. Security Guard

Figure A.10. Truck Driver

Figure A.11. Cleaner

Figure A.12. Carpenter

Figure A.13. Plumber

Figure A.14. Machine Operator

Figure A.15. Technical Support Person

Figure A.16. Software Engineer

Figure A.17. Writer

Appendix B

Debiasing Image-to-Image Translation Models - Supplementary Material

B.1 Train-Validation-Test Split CelebA-HQ

For experimentation with the CelebA-HQ [30] dataset, we make separate datasets for each of the attributes, namely 'Eyeglasses', 'Bald', and 'Wearing Hat.' For fair comparison, we hold out an equal number of images from majority and minority classes for both validation and testing. Note that the number of images from the minority classes are quite low in number (e.g., 700 images for 'Bald'). Therefore, we take 50 images from each class for validation (100 in total) and 150 images for the test set (total 300 images). From the rest of the images, we adjust the training dataset such that the original bias ratio remains unchanged for our selected attributes. The exact number of images for each of the training dataset along with validation and test is given in Table B.1.

B.2 Training Procedure

Data Augmentations. We apply data augmentations to both majority and minority classes. Specifically, these augmentations are shifting 10% (both vertically and horizontally), shearing 10%, scaling 10%, and mirror flipping.

Classifiers. For training, we replace the last fully connected layer of a pretrained

Table B.1. Train-validation-test splits for specific attributes.

Attribute	Group	Train	Validation	Test
Bald	Minority	512	50	150
	Majority	21,062	50	150
	Both	21,574	100	300
Wearing Hat	Minority	870	50	150
	Majority	23,523	50	150
	Both	24,393	100	300
Eyeglasses	Minority	1,268	50	150
	Majority	24,645	50	150
	Both	25,913	100	300

ResNet50 [20], and re-train it again on our training dataset. This classifier is used for applying auxiliary classifier loss, $\mathscr{L}_c$. We train three separate classifiers, one for each attribute. Their performances on the test sets are as follows:

- *Bald.* F1 score: 0.8889, Prediction scores: 87.13%, Accuracy: 90%
- *Eyeglasses.* F1 score: 0.9899, Prediction scores: 95.99%, Accuracy: 99%
- *Wearing Hat.* F1 score: 0.951, Prediction scores: 90.65%, Accuracy: 95.33%

To train the classifier for evaluation, we train a deeper network, ResNet152 [20], on the same training sets. This improves the accuracy for prediction for 'Bald' (94%) and 'Eyeglasses'(99.33%), which makes it reliable for prediction. This also keeps our evaluation classifier network architecture and weights separate from ResNet50 network which was used for $\mathscr{L}_c$ during training. This is critical for fair evaluation as we do not want to evaluate using a classifier which was used for training since the network is optimized to perform well for it. Training the evaluation classifier follows the same procedure. Their performances are:

- *Bald.* F1 score: 0.9362, Prediction scores: 90.16%, Accuracy: 94%
- *Eyeglasses.* F1 score: 0.9933, Prediction scores: 94.67%, Accuracy: 99.33%
- *Wearing Hat.* F1 score: 0.951, Prediction scores: 91.44%, Accuracy: 95.33%

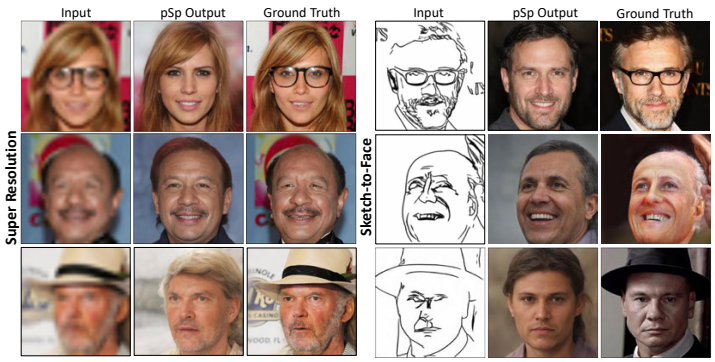


Figure B.1. Example cases of bias for both super-resolution and sketch-to-face. Here, for both tasks, the attributes are visible in the inputs images (i.e. Eyeglasses, Hat, Baldness) but they are missing in the generated images.

Measuring Bias. We use CelebA-HQ [30] to demonstrate how we can measure bias in image-to-image translation tasks. We use CelebA-HQ for two reasons. First, it has 40 labeled binary attributes (e.g. 'Eyeglasses', 'Bald', etc.), making detecting bias easier compared to unlabeled data (such as Flicker-Face-HQ or FFHQ [32]). Second, it is widely used to train both conditional and unconditional generative models [30, 12, 33, 75, 55], making it an ideal dataset for bias analysis. Some additional examples of biases are showed in Figure B.1.

For measuring bias, we generate images for the super-resolution task. We down-sampled the ground truth test images by a factor of 8 and applied the super-resolution pSp network (which has been pre-trained on the same dataset as our classifier). We use a ResNet152 classifier (trained on pSp training set) to calculate the F1 scores on real and generated images for measuring biases. These numbers are reported in Table 1 in the *main* chapter. Note that having a low F1 score means either 1) low recall: the model failed to generate the desired attribute in the generated images (for example, not generating images with eyeglasses where the ground truth images have this attribute), or 2) low precision: the model generated an attribute where it should not have (e.g., producing hair when the person is bald). Both cases represent a bias problem. Therefore, low F1 score on the generated images can reveal the biases for the attribute.

Here, we report the F1 scores on the low resolution inputs and sketches. The numbers appear in Table B.2. The numbers for most of the attributes, especially for 'Eyeglasses' and

Table B.2. F1 scores on low-resolution and sketch input images.

Attribute	Bald	Wearing Hat	Eyeglasses	Blond Hair	Bangs	Black Hair	Male	Heavy Makeup	High Cheekbones	Smiling
Percentage	**2.37**	**3.57**	**4.89**	17.09	18.08	21.97	36.86	45.69	46.16	46.97
Low-Resolution Inputs	0.6488	0.8523	0.6613	0.5891	0.6280	0.8173	0.9414	0.2988	0.4641	0.7595
Sketches	0.2234	0.7215	0.8747	0.0000	0.3366	0.0311	0.7615	0.5590	0.1606	0.5848

'Wearing Hat', show that there is enough information in the input images to generate the attributes in question.

Training. Here we describe our changes to Pixel2Style2Pixel [55] (pSp) and pix2pix. In the case of pSp, it encodes the input images using a feature pyramid backbone [43] and maps them to the extended latent space of a frozen StyleGAN2 [33] generator (pre-trained on FFHQ [32]), $\mathscr{W}+$, which is consists of 18 different 512-dimensional feature vectors, one for each StyleGAN2 layer. We apply our contrastive loss (mentioned in the main chapter), $\mathscr{L}_s$, to the latent codes of each of the layers in $\mathscr{W}+$ separately. The latent codes then are followed by MLP layers, which consists of two linear feed-forward networks with 512 hidden units and a ReLU activation in between (one MLP for each of 18 input layers in StyleGAN2). The temparature parameter in Equation 1 is set to 0.07 for all experiments. Finally, we apply auxiliary classifier loss (Equation 2), $\mathscr{L}_c$, on the outputs of the decoder.

For pix2pix, we apply the U-Net architecture for the sketch-to-image translation model. Similar to pSp, we refer to the original pix2pix model as 'Vanilla' and re-sampling the minority during training as 'Sampling Baseline.' For our model (I+II+III), we make similar changes to pix2pix. Specifically, on top of re-sampling, we apply the supervised contrastive loss, $\mathscr{L}_s$ (Equation 1) to the output of bottleneck layer of the encoder. After applying $\mathscr{L}_s$, we pass the features through MLP layer, ϕ, and add an auxiliary classifier loss (Equation 2), $\mathscr{L}_c$, at the end. We also experiment on other two variations of our model (I+II and I+III).

For training our model for both pSp and pix2pix, we follow a curriculum learning procedure. We introduce our losses (supervised contrastive loss, $\mathscr{L}_s$, and auxiliary classifier loss, $\mathscr{L}_c$) after k iterations. We start with a small value of m for both hyperparameters for supervised contrastive loss, λ_s, and auxiliary classifier loss, λ_c. These hyperparameters are then increased

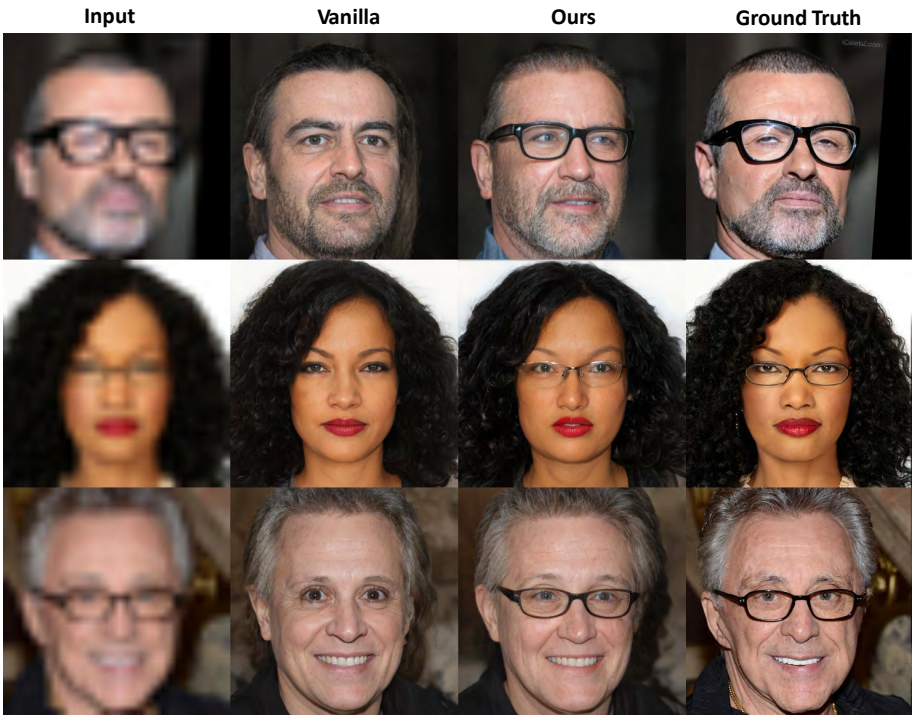


Figure B.2. Results for super-resolution task on 'Eyeglasses'.

by m every k iterations.

For pSp, $k = 10,000$, and $m = 0.001$. We apply similar curriculum training procedure for pix2pix as well. Here, $m = 0.01$. Instead of k iterations, we apply the losses after the first epoch. The values of the hyperparameters are increased by the same value, m, after each epoch. For both pSp and pix2pix, Ours (I+II) and (I+III) follow the same training steps, except the hyperparameter for a specific loss is set to zero. For example, for Ours (I+II), $\lambda_c = 0$, and for Ours (I+III), $\lambda_s = 0$.

B.3 Additional Results and Examples

We show additional examples for the Vanilla and Our Model (I+II+III) from our human face experiments in Figure B.2-B.7.

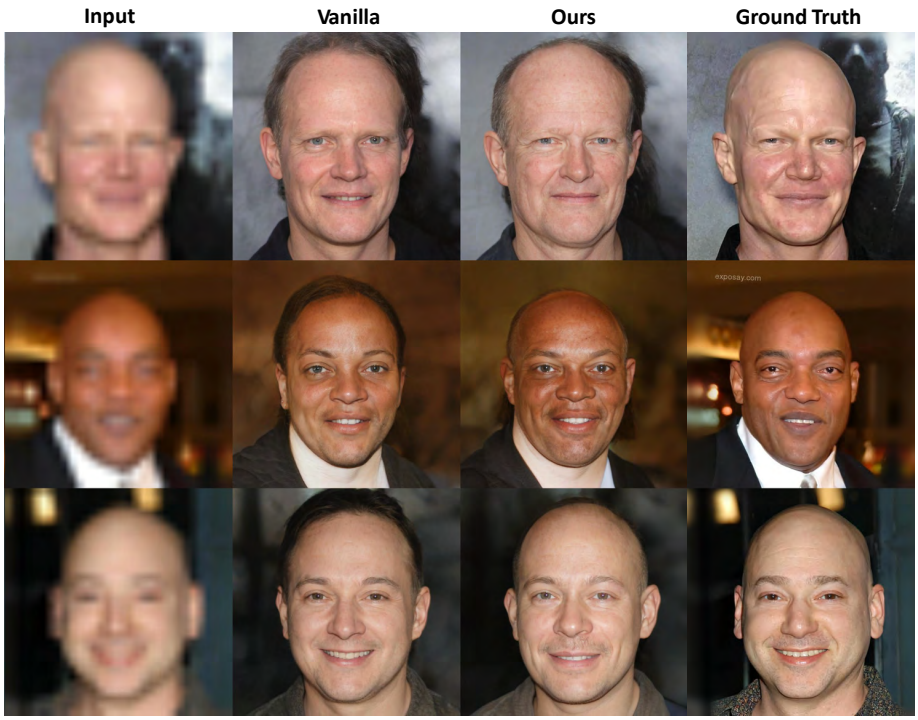


Figure B.3. Results for super-resolution task on ‘Bald’.

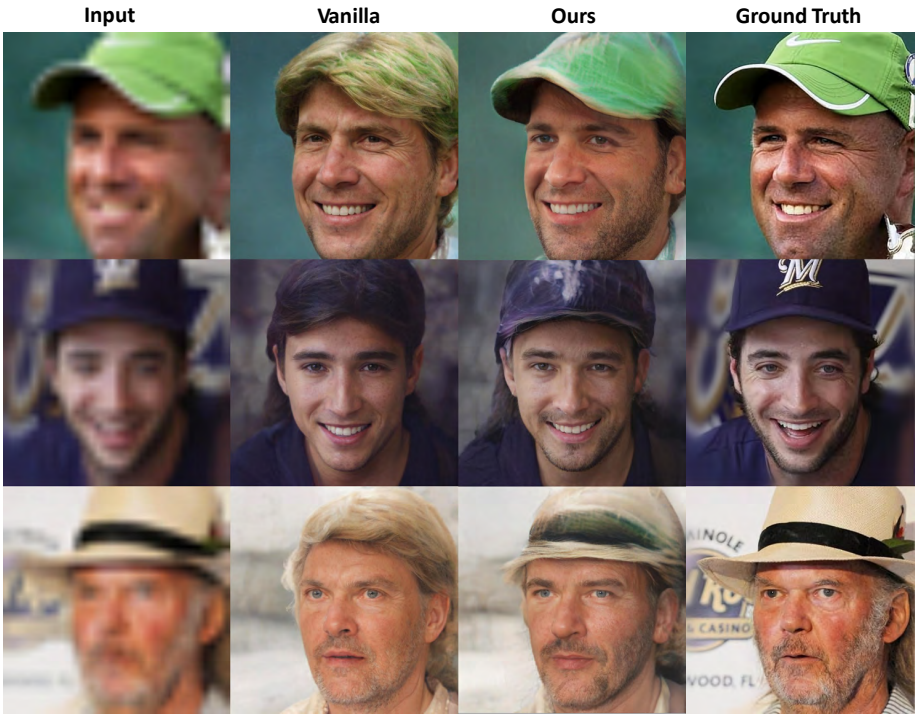


Figure B.4. Results for super-resolution task on ‘Wearing Hat’.

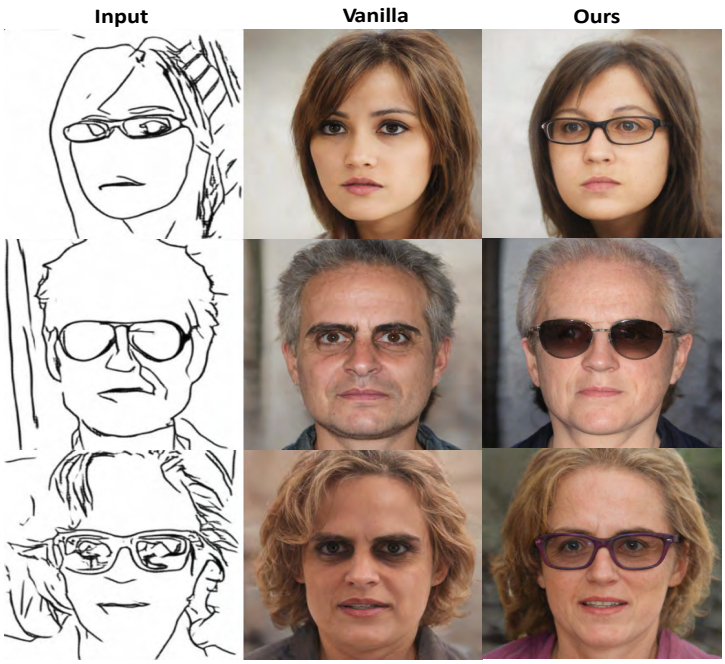


Figure B.5. Results for sketch-to-face task on ‘Eyeglasses’.

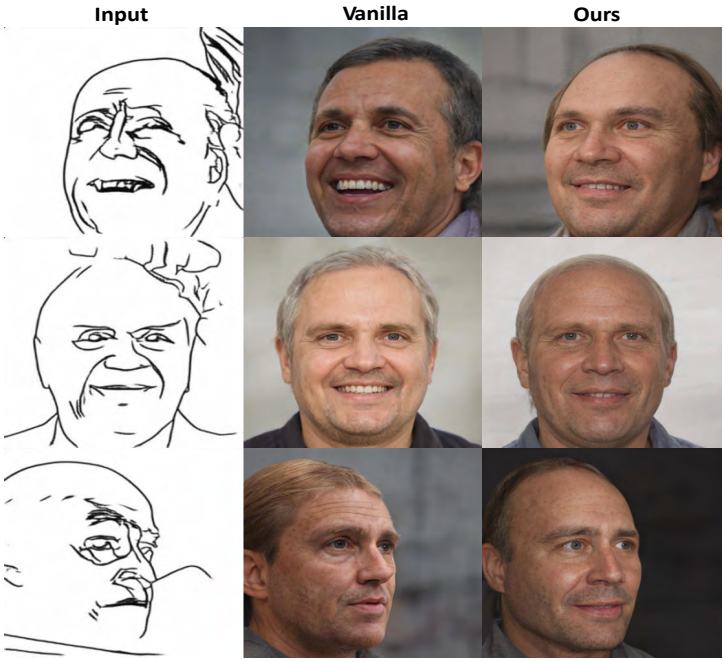


Figure B.6. Results for sketch-to-face task on ‘Bald’.

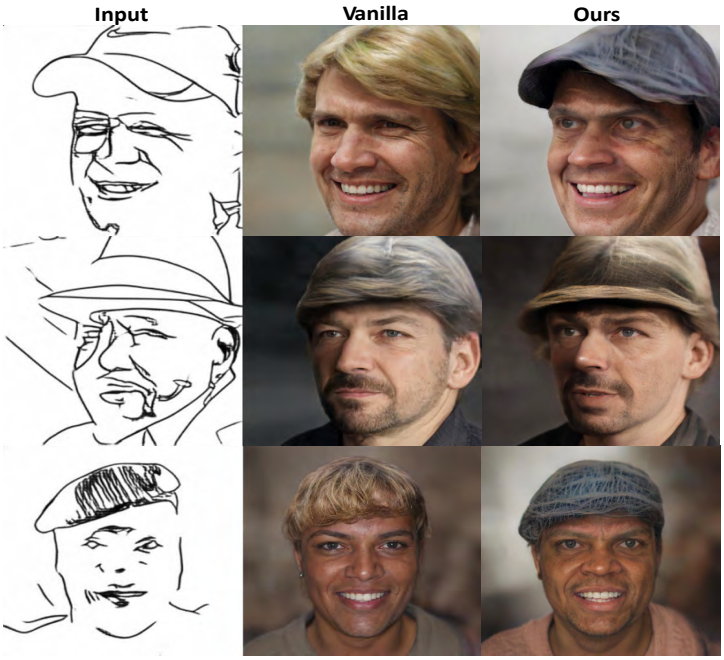


Figure B.7. Results for sketch-to-face task on ‘Wearing Hat’.

Bibliography

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[3] Alexander Amini, Ava P Soleimany, Wilko Schwarting, Sangeeta N Bhatia, and Daniela Rus. Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 289–295, 2019.

[4] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*, pages 327–359. Springer, 2021.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[7] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

[8] L Elisa Celis and Vijay Keswani. Implicit diversity in image summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[9] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. Deepfacedrawing: Deep generation of face images from sketches. *ACM Transactions on Graphics (TOG)*, 39(4):72–1, 2020.

[10] Wengling Chen and James Hays. Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9416–9425, 2018.

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[13] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[14] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[15] Ugur Demir and Gozde Unal. Patch-based image inpainting with generative adversarial networks. *arXiv preprint arXiv:1803.07422*, 2018.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[17] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[18] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27:2672–2680, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[24] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018.

[25] Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020.

[26] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[27] Ajil Jalal, Sushrut Karmalkar, Jessica Hoffmann, Alex Dimakis, and Eric Price. Fairness for image generation with uncertain sensitive attributes. In *International Conference on Machine Learning*, pages 4721–4732. PMLR, 2021.

[28] Chen Karako and Putra Manggala. Using image fairness representations in diversity-based re-ranking for recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 23–28, 2018.

[29] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558, January 2021.

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[31] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[33] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[34] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828, 2015.

[35] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[36] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[37] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[38] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[39] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[40] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

[41] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[42] Jiadong Liang, Wenjie Pei, and Feng Lu. Cpgan: Content-parsing generative adversarial networks for text-to-image synthesis. In *European Conference on Computer Vision*, pages 491–508. Springer, 2020.

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[44] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3673–3682, 2019.

[45] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[46] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. *arXiv preprint arXiv:1910.06809*, 2019.

[47] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

[48] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[49] Jianmo Ni and Julian McAuley. Personalized review generation by expanding phrases and attending on aspect-aware representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 706–711, 2018.

[50] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[51] Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. Elastic-infogan: Unsupervised disentangled representation learning in class-imbalanced data. *arXiv preprint arXiv:1910.01112*, 2019.

[52] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[53] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[55] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[57] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017.

[58] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

[59] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[60] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, and Gerard Medioni. Gan-control: Explicitly controllable gans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14083–14093, 2021.

[61] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11070–11078, 2020.

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

[63] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[64] Md Tanjim, Ritwik Sinha, Krishna Kumar Singh, Sridhar Mahadevan, David Arbour, Moumita Sinha, Garrison W Cottrell, et al. Generating and controlling diversity in image search. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 411–419, 2022.

[65] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[66] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[67] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alexander Graves. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems 29*, pages 4790–4798, 2016.

[68] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[69] Yaxing Wang, Abel Gonzalez-Garcia, Luis Herranz, and Joost van de Weijer. Controlling biases and diversity in diverse image-to-image translation. *Computer Vision and Image Understanding*, 202:103082, 2021.

[70] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[71] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of machine learning research*, 10(2), 2009.

[72] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020.

[73] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pages 776–791. Springer, 2016.

[74] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, et al. Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle). *IEEE transactions on medical imaging*, 39(1):188–203, 2019.

[75] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[76] Ning Yu, Ke Li, Peng Zhou, Jitendra Malik, Larry Davis, and Mario Fritz. Inclusive gan: Improving data and minority coverage in generative models. In *European Conference on Computer Vision*, pages 377–393. Springer, 2020.

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[79] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[80] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5802–5810, 2019.

[81] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.