Generating and Controlling Diversity in Image Search

Md Mehrab Tanjim*
UC San Diego
mtanjim@eng.ucsd.edu

Ritwik Sinha
Adobe Research
risinha@adobe.com

Krishna Kumar Singh
Adobe Research
krishsin@adobe.com

Sridhar Mahadevan
Adobe Research
smahadev@adobe.com

David Arbour
Adobe Research
arbour@adobe.com

Moumita Sinha
Adobe Applied ML
mousinha@adobe.com

Garrison W. Cottrell
UC San Diego
gary@eng.ucsd.edu

Abstract

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 paper, 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.

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 1 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 affects both image search and stock platforms with paying customers.

Figure 1. Top image results retrieved from Google Image search for the query 'plumber' reveal intrinsic biases.[1]

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 [9, 2], 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-

*Work done during an internship at Adobe Research.
[1]Search conducted in January, 2021 from California.
[2]When we first conducted this research. Search engines may have been updated since.

Figure 2. High-resolution images generated for the set of keywords from our proposed model 'Uniform+.'

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) [5], have great potential because of their ability to produce photo-realistic images either unconditionally [13, 14] or conditionally [23, 17, 19, 18].

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, and so on (which is apparent in Figure 1). 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 [13] or StyleGAN2 [14], 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) [7] 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 [7] and classification accuracy (similar to [3]) 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 which show exciting results for combating bias in image search.

2. Related Work and Background

2.1. Bias in Image Search

To characterize the gender bias in image search results for a variety of occupations, the authors of [15] 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 [2] 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, [9] propose a Fairness Maximal Marginal Relevance (FMMR) retrieval algorithm

to reflect diversity in the top image search results. Similar work is explored in [2]. 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. Attribute-to-Image Synthesis Models

In recent years, Generative Adversarial Networks or GANs [5] 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 [21, 19, 22, 1] or text input have been proposed [23, 18, 17]. 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.

2.3. 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 [1], 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 [11], 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 [13] and its improved version StyleGAN2 [14] both combat this entanglement problem by introducing a mapping network and adaptive instance normalization (AdaIN) [8] 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 paper, we overcome these new challenges in our proposed task.

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.

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[19] and text-to-image synthesis generative models such as DMGAN[23], CPGAN[18], ObjGAN[17] 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[13] and StyleGAN2[14], 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.

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[13]: the latent codes $\mathbf{z} \in \mathcal{Z}$ are first transformed to intermediate latent space $\mathbf{w} \in \mathcal{W}$ by a non-linear mapping network $f : \mathcal{Z} \rightarrow \mathcal{W}$. 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

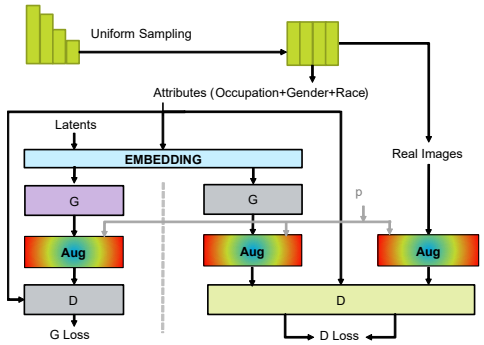

Figure 3. Model architecture and training framework for Uniform+. In addition to class-conditioning with regularization, we introduce new sampling techniques to handle the class imbalance (uniform sampling with augmentation). This component is not specific to StyleGAN2 and can be applied to any generative model.

(AdaIN) [8] after each convolutional layer of the generators of progressive GAN [11]. 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 $\mathcal{W}$ essentially controls styles within convolutional layers at each resolution through AdaIN. It is shown in [13] that these design choices for StyleGAN lead to a less entangled latent space in $\mathcal{W}$ compared to the input latent space in $\mathcal{Z}$. StyleGAN2 [14] 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.

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: $\mathbf{y} = [\mathbf{y_{occupation}}|\mathbf{y_{gender}}|\mathbf{y_{race}}]$. 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 [1] does. Rather, we make changes to the discriminator D_ϕ and apply the zero-centered gradient penalty from [20] 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:

$$D_\phi(\mathbf{x}|\mathbf{y}) = \sum f_\phi(\mathbf{x}) \odot \mathbf{y} \tag{1}$$

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 [5] which is used in [14] for high-quality face generation from their Flickr-Faces-HQ dataset (FFHQ) [13]:

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

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.'

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 [4], 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 (Equation1) which is used in the loss function:

$$D_\phi(\mathbf{x}|\mathbf{y}) = \sum f_\phi(\mathbf{x}) \odot (\mathbf{m} \odot \mathbf{y}) \tag{2}$$

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 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

Attributes	Exec. Mangr.	Admin. Asst.	Nurse	Farmer	Military	Security	Truck	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

Table 1. Data statistics of Stock-Occupation-HQ. The breakdown shows the imbalance in race and gender across different professions.

it starts to increase continuously, and training begins to diverge. Nevertheless, we observed some improvement over Weighted. In the rest of the paper, 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 [12] 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 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.

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.

4.1. Collection

To obtain images for different occupations, we first construct our search query according to [15]. 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).

Figure 4. Real images from each profession after preprocessing

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 1 shows this. To overcome these challenges, we first use dlib's [16] 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 4 shows one example from each job, in the same order as the list above[3].

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 [6] classifier that has been pre-trained on the recently proposed dataset on fairness tasks: FairFace [10]. Its average accuracy is 95.7% for gender and 81.5% for race (Table 8 in [10]). 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 1. As can be seen from the

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

Figure 5. Generated images from the four highest-scoring models show the qualitative differences. ADA achieves the lowest FID but it often fails to generate images faithful to the attributes. While both Vanilla-FFHQ and Uniform do a better job in that , they suffer from low variability and/or artifacts. But the qualitative inspection reveals that Uniform+'s images achieve a better balance in all aspects.

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.

5. Experiments

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 [13] 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 3.2) for all models.

Metrics: For our first metric, we use the popular FID [7] 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 [3]. 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 [10]. For detecting profession we train a ResNet56 [6] on our dataset which achieves 80.28% top-1 accuracy (94.57% top-3). Using them, we compute the AMS for each attribute.

[4]https://github.com/NVlabs/stylegan2

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

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**

Table 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.

5.2. Quantitative Results

Table 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 [12]). 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 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 [12]. 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.

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 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

Figure 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) and thus its high score in some categories (e.g. Occupation AMS) can be misleading. However, this is not the case for Uniform+.

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 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+.

5.3.1 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 matching attributes does not capture the comparative

Figure 7. Curated collection of generated images from Uniform+ (from left to right, top to bottom): 'executive manager', 'administrative assistant', 'nurse', 'farmer', 'military person', 'security guard', 'truck driver', 'cleaner', 'carpenter','plumber', 'machine operator', 'technical support person', 'software engineer', 'writer.' Zoom in for a better view. More examples are given in the supplementary material.

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**

Table 3. Human evaluation results (in percentage). For the last two studies, the percentage is calculated among the considered models.

quality of the images among the models (for example, see Figure 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 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). 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 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 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.

6. Conclusion

In this paper, 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 paper 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.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[2] L Elisa Celis and Vijay Keswani. Implicit Diversity in Image Summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Davis E. King. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[17] 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.

[18] 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.

[19] 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.

[20] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which Training Methods for GANs Do Actually Converge? *arXiv preprint arXiv:1801.04406*, 2018.

[21] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[22] 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.

[23] 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.