

Changing Gender and Race in Image Search Results With Machine Learning

Originally published at <https://www.unite.ai/changing-gender-and-race-in-image-search-results-with-machine-learning/>



A research collaboration between UC San Diego and Adobe  ADBE 0.82% Research has proposed an innovative and proactive solution to the lack of racial and gender diversity in image search results for traditionally WASP-dominated occupations: the use of Generative Adversarial Networks (GANs) to create non-real images of 'biased' professions, where the gender and/or race of the subject is altered.



In this example from the new paper, the researchers have input characteristics for a desired photo that is either not represented in a typical corpus of available image material, or else is represented in an unsuitable way (i.e. sexualized or in an otherwise inappropriate representation). [Source](#)

In a new [paper](#) titled *Generating and Controlling Diversity in Image Search*, the authors suggest that there is a limit to the extent that [re-ranking](#) can fix the imbalance of biased image/feature classes such as *plumber*, *machine operator*, *software engineer*, and many others – and that increasing racial and gender diversity with synthetic data may be the way forward for this challenge.

'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.'

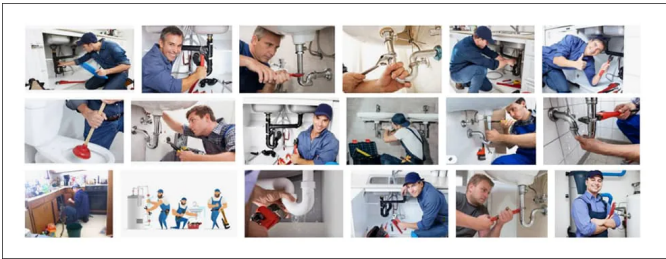
To this end, the authors have experimented with a variety of GAN-based image synthesis systems, finally lighting on an architecture based around [StyleGan2](#).



From the supplementary materials for the paper, two examples of ‘equalizing’ image-based representations of biased professions, in these cases, ‘carpenter’ and ‘

Inadequately or Inappropriately Represented

The researchers frame the challenge in terms of a real-world [search result for ‘plumber’](#)* on Google Image search, observing that image results are dominated by young white men.



From the paper, select results for ‘plumber’ in Google Image search, January 2021.

The authors note that similar indications of bias occur for a range of professions, such as ‘administrative assistant’, ‘cleaner’, and ‘machine operator’, with corresponding biases for age, gender, and race.

‘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].

‘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.’

The paper cites another academic collaboration [from 2014](#), where researchers collected the top 400 image search results for 96 occupations. That work found that women represented only 37% of results, and anti-stereotypical images only 22%. A [2019 study from Yale](#) found that five years had brought these percentages up to only 45% and 30% respectively.

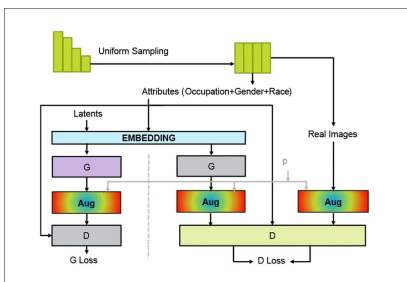
Additionally the 2014 study classified the sexualization of individuals in certain occupations in image search results as the *Sexy Carpenter Problem*, with such inappropriate classifications potentially skewing results for occupation recognition.

The Big Picture

The primary challenge for the authors was in producing a GAN-based image synthesis system capable of outputting 1024×1024 resolution, since, at the current state of the art in GAN and encoder/decoder-based image synthesis systems, 512×512 is pretty luxurious. Anything higher would tend to be obtained by upscaling the final output, at some cost of time and processing resources, and at some risk to the authenticity of the generated images.

However, the authors state that lower resolutions could not expect to gain traction in image search, and experimented with a variety of GAN frameworks that could be capable of outputting hi-res images on demand, at an acceptable level of authenticity.

When the decision was made to adopt StyleGan2, it became apparent that the project would need greater control over sub-features of the generated output (such as race, occupation, and gender), than a default deployment permits. Therefore the authors used multi-class conditioning to augment the generation process.



The architecture of the specifying image generator, which the authors state is not specific to StyleGAN2, but could be applied across a range of generator frameworks.

To control the factors of race, gender, and occupation, the architecture injects a one-shot encode of these concatenated characteristics into the y vector. After this, a feedforward network is used to embed these features, so that they will not be disregarded at generation time.

The authors observe that there are hard limitations to the extent that StyleGAN2 can be manipulated in this way, and that more fine-grained attempts to alter the outcomes resulted in poorer image quality, and even [mode collapse](#).

These remedies, however, do not solve implicit bias problems in the architecture, which the researchers had to address by oversampling under-represented entities from the dataset, but without risking to overfit, which would affect the flexibility of the generated image streams.

Therefore the authors adapted [StyleGAN2-ADA](#), which uses Adaptive Discriminator Augmentation (ADA), to prevent the discriminator from overfitting.

Data Generation and Evaluation

Since the objective of the project is to generate new, synthesized data, the researchers adopted the methodology of the 2014 project, choosing a number of target professions that demonstrate a high racial and gender bias. The professions chosen were 'executive manager', 'administrative assistant', 'nurse', 'farmer', 'military person', 'security guard', 'truck driver', 'cleaner', 'carpenter', 'plumber', 'machine operator', 'technical support person', 'software engineer', and 'writer.'

The authors selected these professions not only based on the extent of perceived bias in image search results, but because most of them contain some kind of visual component that is codified to the profession, such as a uniform, or the presence of specific equipment or environments.

The dataset was fueled by 10,000 images from the Adobe Stock library, typically obtaining a 95% score or better when attempting to classify a profession.

Since many of the images were not helpful for the target task (i.e., they did not contain people), manual filtering was necessary. After this, a [ResNet32](#)-based classifier pretrained on [FairFace](#) was used to label the images for gender and race, obtaining an average accuracy of 95.7% for gender and 81.5% for race. Thus the researchers obtained image labels for the attributes *Sex: Male, Female, Race: White, Black, Asian, and Other Races*.

Models were built in TensorFlow using StyleGAN2 and StyleGAN2-ADA as core networks. Pretraining was done with StyleGAN2's pre-trained weights on the NVIDIA's Flickr-Faces-HQ Dataset ([FFHQ](#)) dataset, augmented with 34,000 occupation-specific images which the authors gathered into a separate dataset that they named Uncurated Stock-Occupation HQ (U-SOHQ).



A sample HIT from the Amazon Mechanical Turk human evaluation.

Images were generated under four configurations of architecture, with Uniform+ finally obtaining the best scores both in FID (automated evaluation), and in subsequent evaluation by Amazon Mechanical Turk workers. Combined with Classification Accuracy, the authors used this as a core metric for their own metric, titled Attribute Matching Score.

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

Human evaluation of images generated by various methods, with the Uniform+ method proving the most convincing, and subsequently the basis for a new dataset.

The paper does not state whether Stock-Occupation-HQ, the full dataset derived from Uniform+, will be made publicly available, but states that it contains 8,113 HQ (1024×1024) images.

Diffusion

The new paper does not explicitly deal with the way that synthesized, 'rebalanced' images could be introduced into circulation. Presumably, seeding new (cost-free) computer vision datasets with redressed images of the type the authors have created would solve the problem of bias, but could also present obstacles to other types of research that seek to evaluate gender and race inclusion in 'real world' scenarios, in a circumstance where synthetic images are mixed with real-world images.

Synthetic databases such as that produced by the researchers could presumably be made available at no cost as reasonably high-resolution stock imagery, using this cost-saving incentive as an engine of diffusion.

The project does not address age-based bias, presumably a potential topic of interest in future research.

* Captured search conducted 5th January 2022, the authors' search cited in the paper was conducted in January of 2021.

First published 5th January 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.

Portfolio site: martinanderson.ai

Contact: martin@martinanderson.ai



Discover More