Are GANs Biased? Evaluating GAN-Generated Facial Images via Crowdsourcing

Hangzhi Guo, Lizhen Zhu, Ting-Hao 'Kenneth' Huang
College of Information Sciences and Technology
Penn State University
{hangz,ljz5180,txh710}@psu.edu

Abstract

Generative models produce astonishingly high-resolution and realistic facial images. However, reliably evaluating the quality of these images remains challenging, not to mention performing a systematic investigation of the potential biases in generative adversarial models (GAN). In this paper, we argue that crowdsourcing can be used to measure the biases in GAN quantitatively. We showcase an investigation that examines whether GAN-generated facial images with darker skin tones are of worse quality. We ask crowd workers to guess whether the image is real or fake, and use this as a proxy metric for estimating the quality of facial images generated by state-of-the-art GANs. The results show preliminary evidence that GANs can generate worse quality images with darker skin tones than images with lighter skin tones. More research is needed to understand the sources, effects, and generalizability of this observed phenomenon.

1 Introduction

The development of image generative models has reached a staggering stage, but the evaluation of GAN cannot catch up with the rapid model developments. In particular, it is commonly overlooked the disparity in image generation quality of different racial groups. Widely adopted automatic evaluation methods [17, 8] commonly measure the quality of generated images by using pre-trained models (*e.g.*, Inception Net [19]) to calculate the similarity of representation of "real" images and "fake" images. As such, these automatic metrics inherent flaws of deep learning models and might fail to reflect to the overfitting problems in the model [2, 3]. On the other hand, humans can serve as a robust evaluator of the quality of image generations. A well-trained expert can detect machine-generated images based on common patterns in generated images. Unfortunately, the time for computer-vision experts is limited, and it is infeasible to ask them to devote all their hours to detecting "fake" patterns in machine-generated images. Thus, for the scalability considerations, we propose to *leverage crowdsourcing to evaluate the quality of image generations*.

This paper uses crowdsourcing to evaluate biases in the face generations' quality. (In other words, we ask the crowd workers to play a role similar to the *discriminative network* in GAN.) Although distinguishing real/fake images is seemly challenging for average workers, we show that under some training with some pre-set rules, crowd workers are capable of distinguishing real and fake facial images. By leveraging this framework, we expose the fairness issues on StyleGAN V2 [10], as we observed preliminary evidence of the better quality of lighter skin faces than darker skin faces.

2 Related Works

Evaluating Generated Images. Automatic metrics focus on evaluating the *quality* and *diversity* of generated images. Inception Score [17] and Fréchet Inception Distance [8] are two practical and

NeurIPS 2022 Workshop on Human Evaluation of Generative Models.

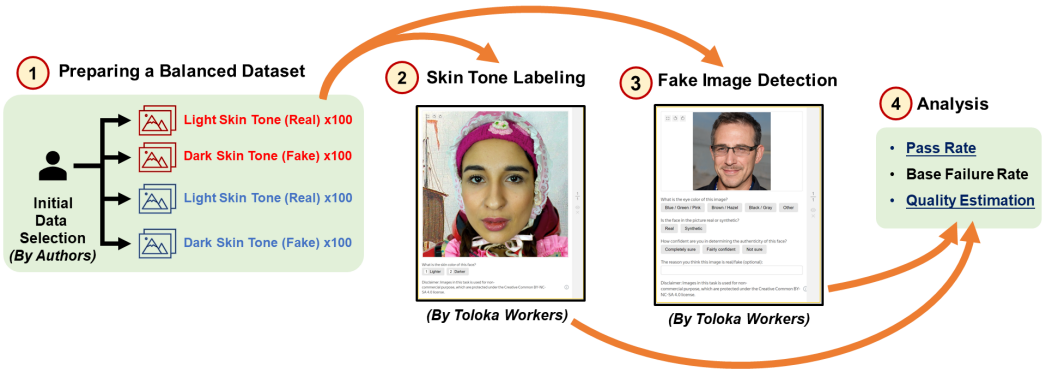

Figure 1: Overview of the evaluation pipeline. We first create a dataset for crowd workers to identify the real and fake facial images (Step 1). We then ask crowd workers to label the skin tone of each face (Step 2) and guess whether the image was a real or fake image (Step 3). Finally, we use the ratio of GAN-generated faces that pass the human examination to evaluate the image quality (Step 4).

popular evaluation methods. These two metrics are based on the classifier Inception Net-V3 [19]. The inception score evaluates the fidelity and diversity of generated images. It is an indirect quality evaluation method that heavily relies on the classifier. Fréchet Inception Distance calculates the Fréchet distance between true and generated data distributions. Its limitations lie in not being able to reflect overfitting and relying on the features learned by the classifier. Meanwhile, human perception metrics can serve as complementary evaluations to automatic metrics. In particular, crowdsourcing platforms are often used to measure human perception. Zhou et al. [20] constructed Human eYe Perceptual Evaluation (HYPE) that directly assesses the quality of generative models. $HYPE_{time}$ explores and rates on the minimum time threshold for workers to consistently judge the authenticity of the images. $HYPE_{\infty}$ gives workers unlimited time to judge the authenticity of an image and reflects the quality of the image by workers' accuracy rate.

Biases in ML Models and Applications. Bias is pervasive in machine learning models [14]. Bias arises during the data curation process [18, 15, 1], which is the primary source of biased outcomes from ML models [11, 6]. In addition, algorithmic design choices (e.g., optimization functions, regularization) can also contribute to biased decision outcomes [5]. Biases for people with different attributes can cause serious ethical concerns and severe negative social impacts in the real-world application of models. Lambrecht and Tucker [12] reported discriminative behaviors against females of COMPAS in promoting jobs in STEM fields. Buolamwini and Gebru [4] reported that the accuracies of women and dark people are significantly lower than those of men and light people in various popular facial recognition products. However, these results are detected in the supervised classification models, whereas the biases in the generative models are often overlooked. To fill this gap, our study measures the biases of image generation quality by leveraging crowdsourcing.

3 Method

This paper studies whether GAN-generated facial images of different skin tones are of the same quality. We used *how easily humans can detect a fake face image* as a quality indicator. The central hypothesis is that the worse-generated facial images are easier to detect. Figure 1 overviews this evaluation procedure, which we elaborate on in this section.

Step 1: Preparing a Dataset of Facial Images. We first created a dataset for crowd workers to identify real and fake facial images. The resulting dataset contains 200 real facial images and 200 GAN-generated (fake) facial images. To generate fake facial images, we use StyleGAN V2 [10], as this is one of the state-of-the-art generative models in the image generation task. StyleGAN advances the state of the art by proposing a novel generator design to control the style of the image generation process [9]. The updated V2 model [10] regularizes the generator to further improve realistic mapping to images. To select real images, we sample real images from Flickr-Faces-HQ Dataset [9], which was subsequently used to train the StyleGAN model.

Given the imbalanced skin tones in the Flickr-Faces-HQ datasets (*i.e.*, darker skin images consist of 10% total images), to ensure fairly balanced datasets seen by workers, the authors pre-selected images to ensure the same amount of data in each group (*i.e.*, real or fake images with lighter or darker skin tone; see Figure 1). Next, we ask workers to label the skin tones of these pre-selected facial images (see Step 2). Importantly, we use crowd workers' aggregated labels (instead of authors' labels) as each image is labeled by only one author.

Step 2: Labeling Facial Images' Skin Tones via Crowdsourcing. Following practices in Raji et al. [16], we label the skin tone by using Fitzpatrick skin type of 1 to 3 as "Lighter" skin, and 4 to 6 as "darker" skin tone. To facilitate correct skin tone labeling, we provide an annotated instruction on which skin tone should be labeled as *Lighter* and *Darker*. This annotated instruction contains examples of each skin type. In addition, we require workers to complete a training session to learn the scale of skin tones. Workers are only qualified to label skin tones after they passed the training session. During the labeling process, workers see a simple multi-choice question (*i.e.*, What is the skin color of this face?). Finally, we implement majority voting to aggregate the final results.

Step 3: Fake Face Image Detection Using Crowdsourcing. We leverage crowd workers to evaluate the quality of AI-generated images. We ask workers to detect fake facial images from a mixed real and fake set (collected from Step 1). For each task, we show crowd workers a facial image, and ask workers to guess whether the image is *real* or *fake*.

To facilitate workers excel in this task, we summarize some common patterns of detecting fake images to help workers in detecting fake facial images. Although the generated face images are increasingly realistic, we can find some common patterns to detect the fake facial images [7]. For instance, [13] summarizes some unique characteristics of generated images, *e.g.*, some of them have a surreal background, asymmetry, weird teeth, messy hair. Following these rules, we wrote an instruction to facilitate workers in detecting fake images (see Figure 5 in the Appendix).

Training Sessions. Similar to labeling the skin tones, we require workers to complete a training session. We carefully select 10 real faces, and 10 fake faces. Each fake face can be detected using the rules in our instructions (see Figure 5). During the training session, when workers fail to detect a fake image, we will provide explanations on how we can detect these fake images using our instructions.

Attention Questions. To improve label quality, we additionally collect attention questions during the detection task. We ask workers about the eye color of the detecting images (and variants of questions). These questions intend to maximize workers' attention in detecting the images,

Step 4: Data Analysis. Finally, we use the ratio of GAN-generated faces that pass the human examination as a proxy metric for evaluating image quality. We calculate the *pass rate* and then adjust it using the *base failure rate* to derive the final *quality estimation* (see Figure 2).

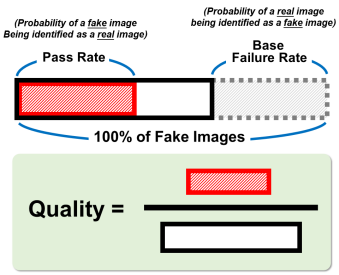

Figure 2: We use how often the GAN-generated faces could pass the human examination as a proxy metric for evaluating the image quality. We will first calculate the **pass rate** and then adjust it using the **base failure rate** to get the final quality estimation.

Pass Rate shows how often a GAN-generated image passes humans who deem it to be a real image.

$$Pass\,Rate = \frac{|\{\text{fake images}\} \setminus \{\text{predicted real images}\}|}{|\{\text{fake images}\}|}$$

Base Failure Rate indicates how often a *real* image fails to pass human judgment and is viewed as faked. It represents the minimum chance that an image could be considered fake. Namely, even when the GAN model produces perfectly-indistinguishable images to real images, each GAN-generated image still has this much chance of being viewed as fake.

$$Base\,Failure\,Rate = \frac{|\{\text{real images}\} \setminus \{\text{predicted fake images}\}|}{|\{\text{real images}\}|}$$

Quality is estimated by adjusting the pass rate to account for the fact that even the perfectly-real images can sometimes be viewed as fake. Figure 2 overviews the concept.

$$Quality = \frac{Pass\ Rate}{1 \quad Base\ Failure\ Rate}$$

We are aware that, besides the image generation model, workers could also be a source of biases. The Base Failure Rate used in our calculation and the use of a balanced dataset (instead of an imbalanced one) aimed to mitigate workers' biases, but eliminating biases in people's judgment is impossible.

4 Experimental Results

Experimental Setups. We conducted the study using Toloka AI.[1] We asked workers to label 400 images for skin-tone labeling and fake/real classification. For each image, the labeling of skin tones is done by 3 workers (*i.e.*, 1,200 labels in total), and the classification of fake/real images involves 5 workers (*i.e.*, 2,000 labels in total). Workers qualify to each task by achieving at least 60% score.

Hourly Wages for Crowd Workers. According to Toloka's guidelines for wages,[2] the average minimum wage in the countries Toloka features is $1.04 per hour. Toloka's pre-set survey prices are about 1.5 times higher. As this hourly wage is significantly lower than the minimum wage in the U.S. (*i.e.*, $7.25), where the authors' institute is located, we decided to conduct two studies. Study 1 followed Toloka's suggested hourly wage, and Study 2 replicated Study 1 but with a higher pay rate.

Skin Tone Labeling Results. Figure 3 shows the distribution of skin tones in our dataset, labeled by the crowd. In particular, 64% (256/400) of the facial images are labeled as lighter skin tones by crowd workers. The fake image pool contains lighter images than the real image pool (133 vs. 123, respectively.) Note that from the authors' labels, the lighter and darker faces have the same ratio. We analyze this disagreement in skin tone labels in Appendix A.

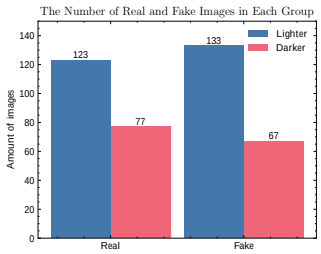

Figure 3: The amount of lighter and darker images labeled by crowd workers.

Automatic Image Quality Evaluation. Table 1 shows the Inception Score (IS), a pervasive automatic metric to evaluate the quality of image generation in each image pool. Notably, fake darker images achieve a lower average IS than fake lighter images, which indicates that the quality is worse of GAN in generating faces with darker skin tones than lighter ones. However, this result is not statistically significant (p-value > 0.05)

Table 1: Inception score for each group. Higher Inception Score indicates better-generated image quality.

Image Pools	Inception Score	
Real-Lighter	3.693	0.693
Fake-Lighter	2.598	0.240
Real-Darker	3.358	0.549
Fake-Darker	2.423	0.369

Table 2: Experimental results of Study 1 (Price = $0.01).

Skin Tone	Att.	Pass Rate		Quality	
Darker	**No Att.**	0.287	0.247	0.402	0.347
	Att.	0.409	0.271	0.608	0.402
Lighter	**No. Att.**	0.391	0.282	0.549	0.396
	Att.	0.402	0.278	0.597	0.413

4.1 Study 1: Fake Face Identification Task with a Regular Pay Rate ($0.01)

In Study 1, we followed Toloka's suggested hourly wage. Based on our estimated task duration, a skin-tone labeling task was priced at $0.01, and a fake/real labeling task was priced at $0.01.

Fake Face Identification Task by Workers. Table 2 shows the results of the experiment with and without attention questions (*i.e.*, the eye color question). In particular, the experiment *without* attention questions shows that both the Pass Rate and Quality of darker images were lower than lighter images, and this gap is *statistically significant* (p-value=0.011 < 0.05). This result indicates that

[1]Toloka AI: an online crowdsourcing platform. https://toloka.ai/

[2]Toloka for Social Sciences: https://toloka.ai/toloka-for-social-sciences/

generated darker faces have poorer quality than lighter faces. On the other hand, when experimenting *with* attention questions, the performance gap (albeit persists) is surprisingly eliminated, as no statistically significant gap exists between different image pools. This result indicates that with attention questions, workers consider the generated lighter and darker images to be of equal quality.

4.2 Study 2: Fake Face Identification Task with a Higher Pay Rate ($0.03)

Study 2 replicated Study 1 with the same parameters and settings, except we used a higher pay rate (*i.e.*, 3X higher than Study 1.) A skin-tone labeling task was priced at $0.03, and a fake/real labeling task was priced at $0.03.

Table 3: Experimental results of Study 2. (Price = $0.03)

Skin Tone	Att.	Pass Rate		Quality	
Darker	**No Att.**	0.328	0.267	0.464	0.377
	Att.	0.367	0.299	0.479	0.389
Lighter	**No. Att.**	0.340	0.275	0.480	0.389
	Att.	0.403	0.294	0.525	0.384

Echoing the results of Study 1. Table 3 highlights the results of Study 2 with and without attention questions under the $0.03 pay rate. Study 2 showed that darker images had lower Pass Rate and Quality measure than lighter images, reiterating the finding in Study 1 that generated darker images have poorer quality than lighter images. However, unlike Study 1, the result of the two experiments in Study 2 is not statistically significant ($p > 0.05$).

The higher wage lowered the image's quality estimation. From Table 2 and 3, we observe that increasing the money rate lowers the quality estimations. One possible explanation could be that the abnormally high wage (*i.e.*, three times the suggested wage on Toloka) motivated workers to pay much more attention to the images. Thus fewer images passed human judgment.

4.3 Worker's Uncertainty Level and Feedback

Workers were often certain about their guesses. To aggregate workers' uncertainty levels, we used 0 for "Completely Sure", 1 for "Fairly Sure", and 2 for "Not Sure" (*i.e.*, the higher the score, the more uncertain the worker is). Table 4 shows that workers were generally certain about their selections.

Worker Feedback. We collected optional free-text feedback from workers. Table 5 shows the number of feedback we received in each setting, alongside their fake/real detection accuracies. Table 6 shows examples of workers' feedback. In particular, workers' feedback highlights inauthentic parts of the image (as perceived by humans), and the feedback for real images often emphasizes some parts perceived as normal by humans. In addition, the accuracy with feedback is higher than the overall accuracy *without attention questions*. On the other hand, the accuracy with feedback is lower *with attention questions*. These results (including results in Table 2) spur future studies in understanding the impact of attention questions on labeling tasks.

Table 4: Workers' uncertainty level score (from 0 to 2) when detecting the fake images. The higher the score, the more uncertain the worker is. In general, workers were quite certain about their selections.

	Real-Lighter		Fake-Lighter		Real-Darker		Fake-Darker	
No Att. ($0.01, Study 1)	0:459	0:621	0:456	0:620	0:436	0:622	0:421	0:585
Att. ($0.01, Study 1)	0:410	0:580	0:402	0:563	0:408	0:580	0:367	0:558
No Att. ($0.03, Study 2)	0:434	0:616	0:408	0:626	0:473	0:641	0:391	0:640
Att. ($0.03, Study 2)	0:328	0:563	0:356	0:563	0:351	0:571	0:334	0:560

Table 5: Statistical results of the workers' feedback (FB).

	#FB	Acc. w/ FB	Overall Acc.
No Att.	125	0.768	0.678
Att.	56	0.500	0.634

Table 6: Examples of workers' feedback.

Feedback	Guess
the design on the shirt is not uniform	Fake
Left side cloths merged in background.	Fake
Image has good detail	Real
Blotch in the right eye	Fake
The face is normal	Real

Acknowledgments

This project was partially supported by Toloka's Research Award. We thank the crowd workers for participating in this project. We also thank the anonymous reviewers for their constructive feedback.

References

[1] Barbosa, S., Cosley, D., Sharma, A., and Cesar Jr, R. M. (2016). Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th International Conference on World Wide Web*, pages 829–841.

[2] Barratt, S. and Sharma, R. (2018). A note on the inception score. *arXiv preprint arXiv:1801.01973*.

[3] Borji, A. (2019). Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65.

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

[5] Danks, D. and London, A. J. (2017). Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697.

[6] Georgopoulos, M., Oldfield, J., Nicolaou, M. A., Panagakis, Y., and Pantic, M. (2021). Mitigating demographic bias in facial datasets with style-based multi-attribute transfer. *International Journal of Computer Vision*, 129(7):2288–2307.

[7] Guo, H., Hu, S., Wang, X., Chang, M.-C., and Lyu, S. (2021). Eyes tell all: Irregular pupil shapes reveal gan-generated faces. *arXiv preprint arXiv:2109.00162*.

[8] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

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

[10] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

[11] Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., and Vetter, T. (2019). Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

[12] Lambrecht, A. and Tucker, C. (2019). Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7):2966–2981.

[13] McDonald, K. (2018). How to recognize fake ai-generated images.

[14] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

[15] Mustard, D. B. (2003). Reexamining criminal behavior: the importance of omitted variable bias. *Review of Economics and Statistics*, 85(1):205–211.

[16] Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., and Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151.

[17] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

[18] Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

[19] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[20] Zhou, S., Gordon, M., Krishna, R., Narcomey, A., Fei-Fei, L. F., and Bernstein, M. (2019). Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32.

A Discussion

Disagreements in Skin Tone Labels. There were 66 (out of 400) disagreements between the authors of this paper and the crowd workers. We analyzed the disagreements and found that many disagreements occurred in the boundary cases, *i.e.*, images' Fitzpatrick skin types of III or IV are hard to label as lighter or darker skin. In addition, the variation in the lighting of a facial image often made the labeling task harder.

How about using different attention questions? In most of our studies, we used a factoid question, *i.e.*, eye color, as the attention question. Although it is a natural attention question as it forces workers to pay close attention to the details in the facial images, one might also use different types of questions in the study. For further exploration, we conducted a small study using *gender* as the attention question (pay rate = \$0.01). In this study, the Quality of darker images is 0.496; the Quality of lighter images is 0.592. Interestingly, we observe that the Pass Rate of darker facial images is lower than lighter facial images, and the discrepancy is statistically significant (p-value $= 0.021 < 0.05$). This result contradicts Study 1, suggesting that workers might behave differently when using subjective or stereotypical questions as the attention question. Future research is needed to understand the impact of attention questions on labeling tasks.

B Worker Instructions

Here, we present instructions to workers in Step 2 and Step 3 (in Figure 1) in our overall pipeline.

Instructions

In this task, you will see images of different faces.
You need to determine their **skin color.**
(Skip the tasks if the image doesn't load or loads only partly.)

How to determine the skin color

Here is a table for measuring the skin colors of faces. Type I, II, and III are considered lighter skin. Type IV, V, and VI are considered darker skin.

Score	Description	Female	Male
0–6 Type I	**Pale white skin** Extremely sensitive skin, always burns, never tans *Example: red hair with freckles*		
7–13 Type II	**White skin** Very sensitive skin, burns easily, tans minimally *Example: fair skinned, fair haired Caucasians, northern Asians*		
14–20 Type III	**Light brown skin** Sensitive skin, sometimes burns, slowly tans to light brown *Example: darker Caucasians, some Asians*		
21–27 Type IV	**Moderate brown skin** Mildly sensitive, burns minimally, always tans to moderate brown *Example: Mediterranean and Middle Eastern Caucasians, southern Asians*		
28–34 Type V	**Dark brown skin** Resistant skin, rarely burns, tans well *Example: some Hispanics, some Africans*		
35+ Type VI	**Deeply pigmented dark brown to black skin** Very resistant skin, never burns, deeply pigmented *Example: darker Africans, Indigenous Australians*		

Image credit: Finding Your Fitzpatrick Skin Type - Injectables

Close

Figure 4: Instructions for labeling skin tones.