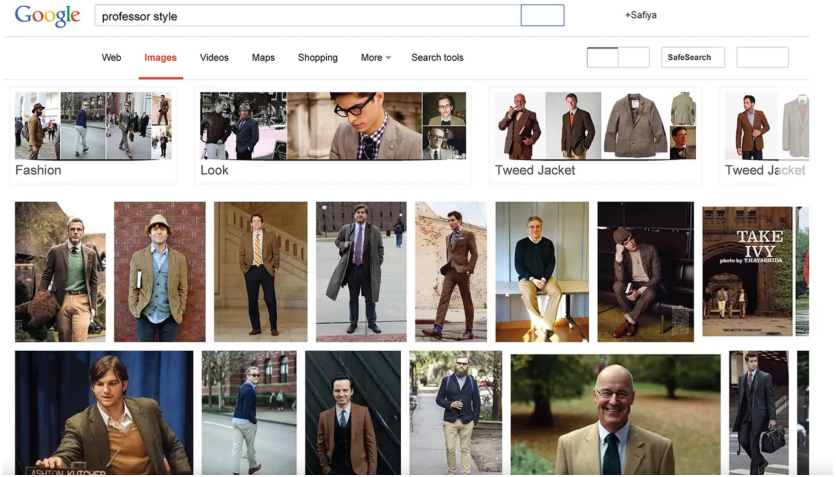
Subverting Fair Image Search with Generative Adversarial Perturbations

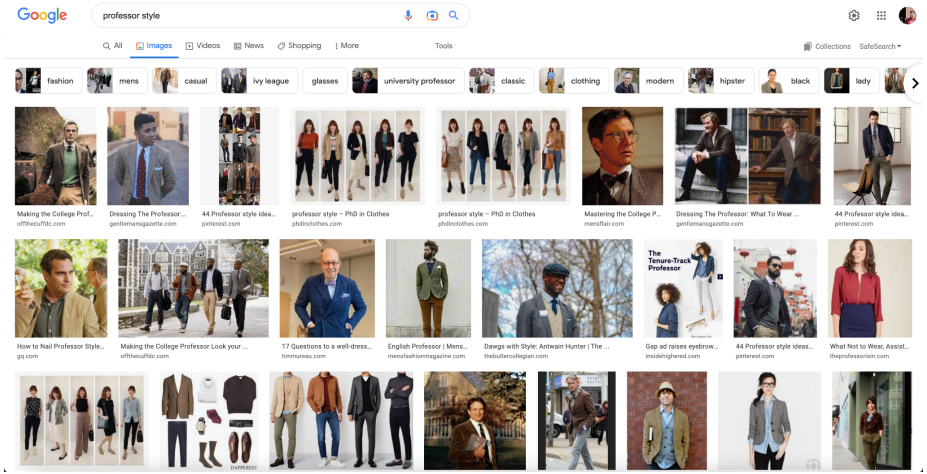
FAccT 2022

Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University, Boston, MA

Fair Image Search



Biased (2015)



Diverse (2022)

Fairness is important for image search. A real-world image search system not only has to crawl images from the internet but should also ideally present diverse, nuanced perspectives.

Fair Ranking

Fair Ranking Algorithms

Several **fair ranking algorithms** have been proposed in the **literature**. Approaches include:

- Constrained optimization (utility/exposure constraint)
- Pairwise comparisons
- Learning-to-rank (amortized fairness under constraints)

Screenshot from Celis et Al, 2018

Fair Ranking Algorithms

Several **fair ranking algorithms** have been proposed in the **literature**. Approaches include:

- Constrained optimization (utility/exposure constraint)
- Pairwise comparisons
- Learning-to-rank (amortized fairness under constraints)

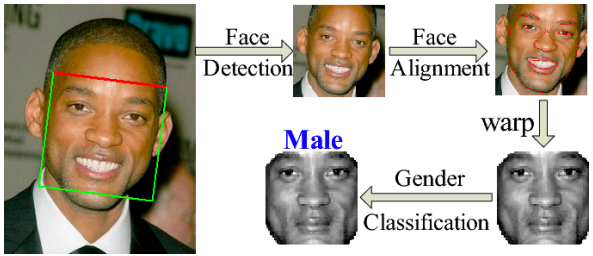
Majority of these algorithms require protected attribute information!

Screenshot from Celis et Al, 2018

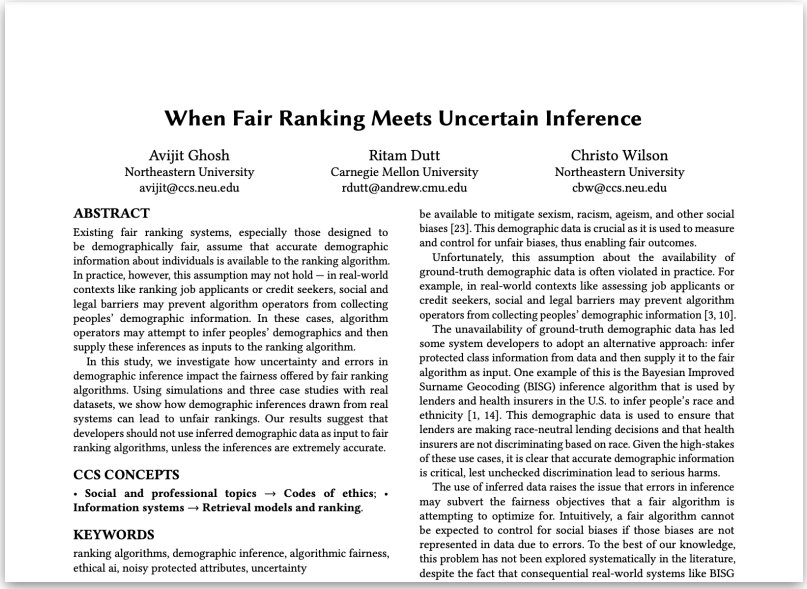
Demographic Classification

Unfortunately, a common workaround is to use **demographic classifiers** that infer the race/gender or other sensitive attribute from people's name, image, zip code, or other information.

Examples: BISG (name, zip) or Deepface (image)



Previous Work



When Fair Ranking Meets Uncertain Inference

Avijit Ghosh
Northeastern University
avijit@ccs.neu.edu

Ritam Dutt
Carnegie Mellon University
rdutt@andrew.cmu.edu

Christo Wilson
Northeastern University
cbw@ccs.neu.edu

ABSTRACT

Existing fair ranking systems, especially those designed to be demographically fair, assume that accurate demographic information about individuals is available to the ranking algorithm. In practice, however, this assumption may not hold — in real-world contexts like ranking job applicants or credit seekers, social and legal barriers may prevent algorithm operators from collecting peoples' demographic information. In these cases, algorithm operators may attempt to infer peoples' demographics and then supply these inferences as inputs to the ranking algorithm.

In this study, we investigate how uncertainty and errors in demographic inference impact the fairness offered by fair ranking algorithms. Using simulations and three case studies with real datasets, we show how demographic inferences drawn from real systems can lead to unfair rankings. Our results suggest that developers should not use inferred demographic data as input to fair ranking algorithms, unless the inferences are extremely accurate.

CCS CONCEPTS

• **Social and professional topics** → **Codes of ethics**; • **Information systems** → **Retrieval models and ranking**.

KEYWORDS

ranking algorithms, demographic inference, algorithmic fairness, ethical ai, noisy protected attributes, uncertainty

be available to mitigate sexism, racism, ageism, and other social biases [23]. This demographic data is crucial as it is used to measure and control for unfair biases, thus enabling fair outcomes.

Unfortunately, this assumption about the availability of ground-truth demographic data is often violated in practice. For example, in real-world contexts like assessing job applicants or credit seekers, social and legal barriers may prevent algorithm operators from collecting peoples' demographic information [3, 10].

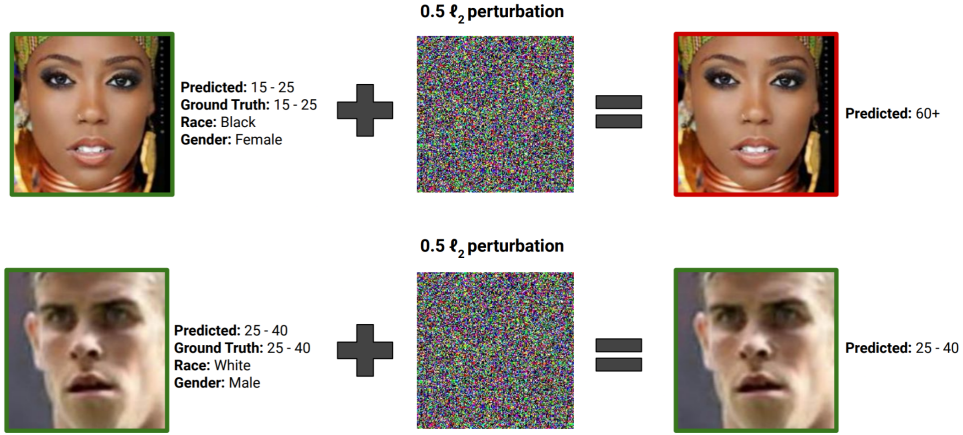
The unavailability of ground-truth demographic data has led some system developers to adopt an alternative approach: infer protected class information from data and then supply it to the fair algorithm as input. One example of this is the Bayesian Improved Surname Geocoding (BISG) inference algorithm that is used by lenders and health insurers in the U.S. to infer people's race and ethnicity [1, 14]. This demographic data is used to ensure that lenders are making race-neutral lending decisions and that health insurers are not discriminating based on race. Given the high-stakes of these use cases, it is clear that accurate demographic information is critical, lest unchecked discrimination lead to serious harms.

The use of inferred data raises the issue that errors in inference may subvert the fairness objectives that a fair algorithm is attempting to optimize for. Intuitively, a fair algorithm cannot be expected to control for social biases if those biases are not represented in data due to errors. To the best of our knowledge, this problem has not been explored systematically in the literature, despite the fact that consequential real-world systems like BISG

- Fair Ranking methods which **require access to demographic information** are prone to **violate fairness guarantees** if this information is **noisy.**
- Sometimes **protected groups can be worse off than rankings without any intervention.**
- The violation is **not easy to predict** and the relationship between per class prediction accuracy and overall effect is **complex.**

Adversarial Attacks can cause bias

Intentional Biases



What if demographic information is untrustworthy because someone is ***intentionally attempting to misrepresent themselves or their data***?

It is possible for the same adversarial perturbation to cause completely different outcomes for people in different subgroups. (Nanda et al. 2021)

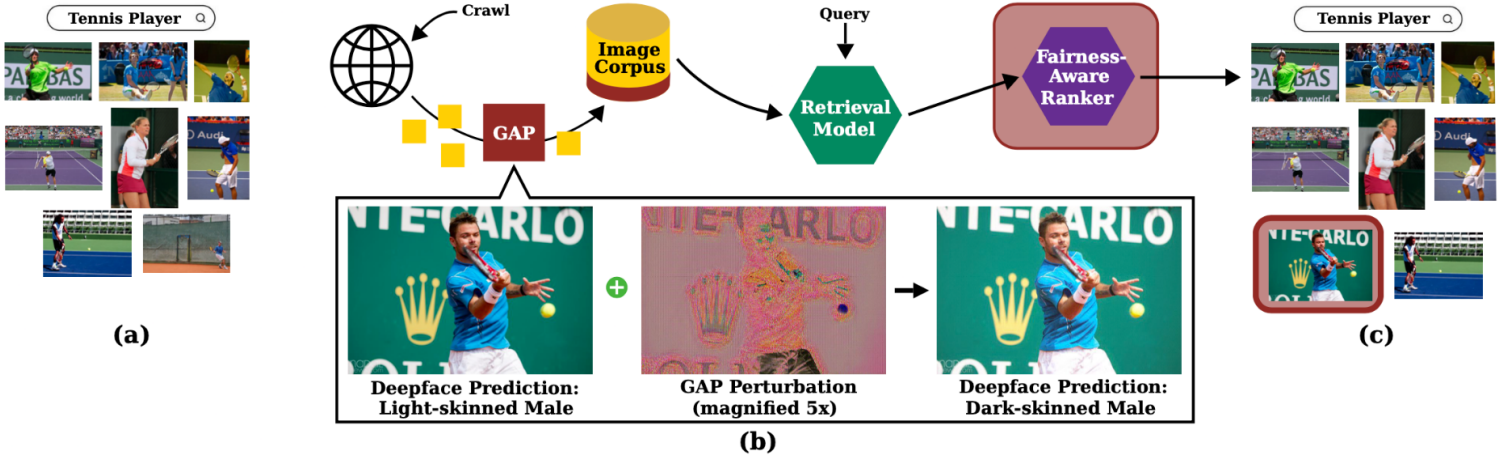
Threat Model

We attack a **Fair Image Search** model that consists of two parts - a **retrieval step and a fair re-ranking step**. The fair re-ranking step either uses inferred labels or image embeddings.

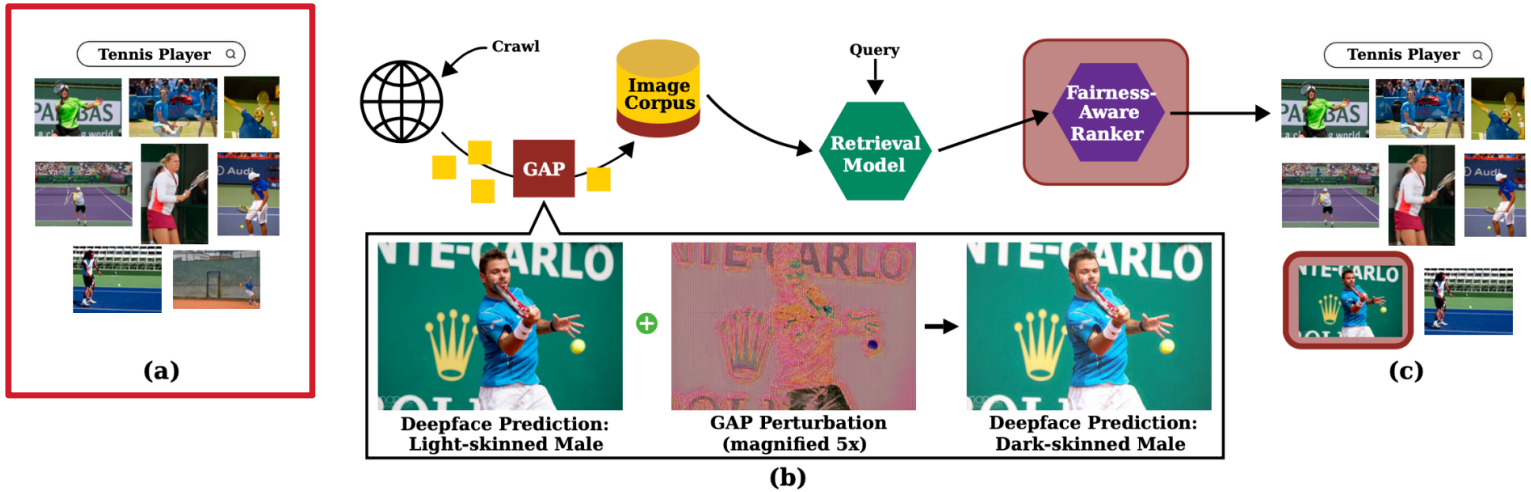
A **very restrictive threat model,** to be as close to a real-world attack as possible.

1. This is an **evasion attack**, which means that the attacker does not have influence on the model parameters.
2. The attacker also does not know which fair-reranking model is being used by the image search system.
3. The attacker uploads adversarially perturbed images onto the internet, and a web scraper collects these images along with other clean images from all over the web and adds to a repository of images to retrieve from.
4. Threat model is similar to Clean Label attacks, or Cloaking Defenses (Shan and Wenger et al, 2020)

Threat Model: A Schematic

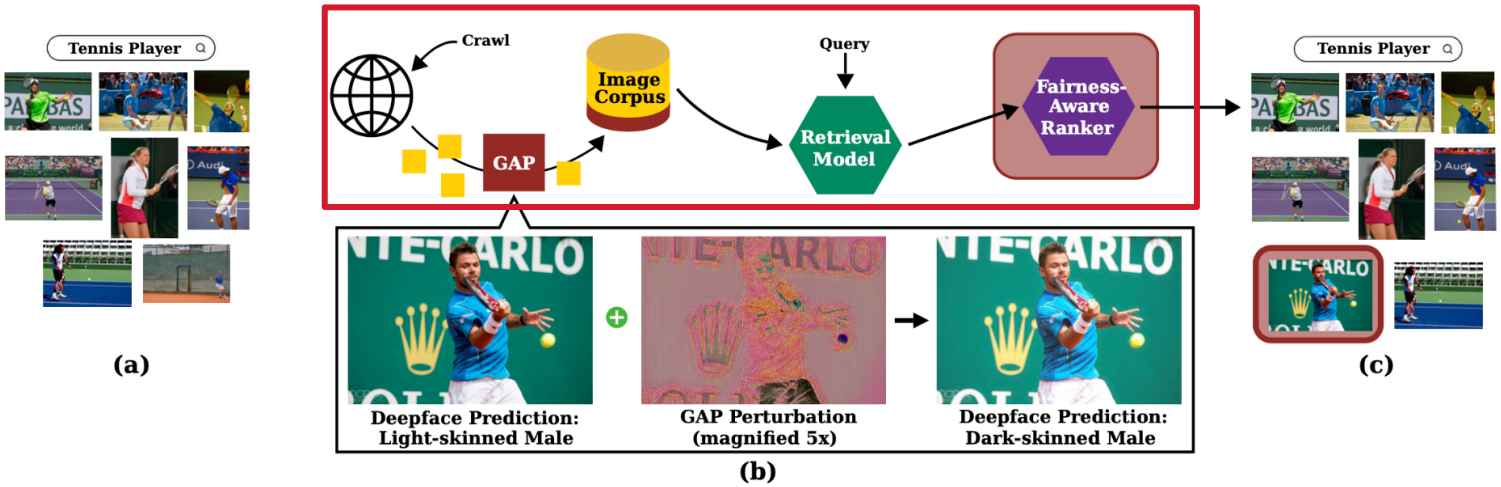


Threat Model: A Schematic



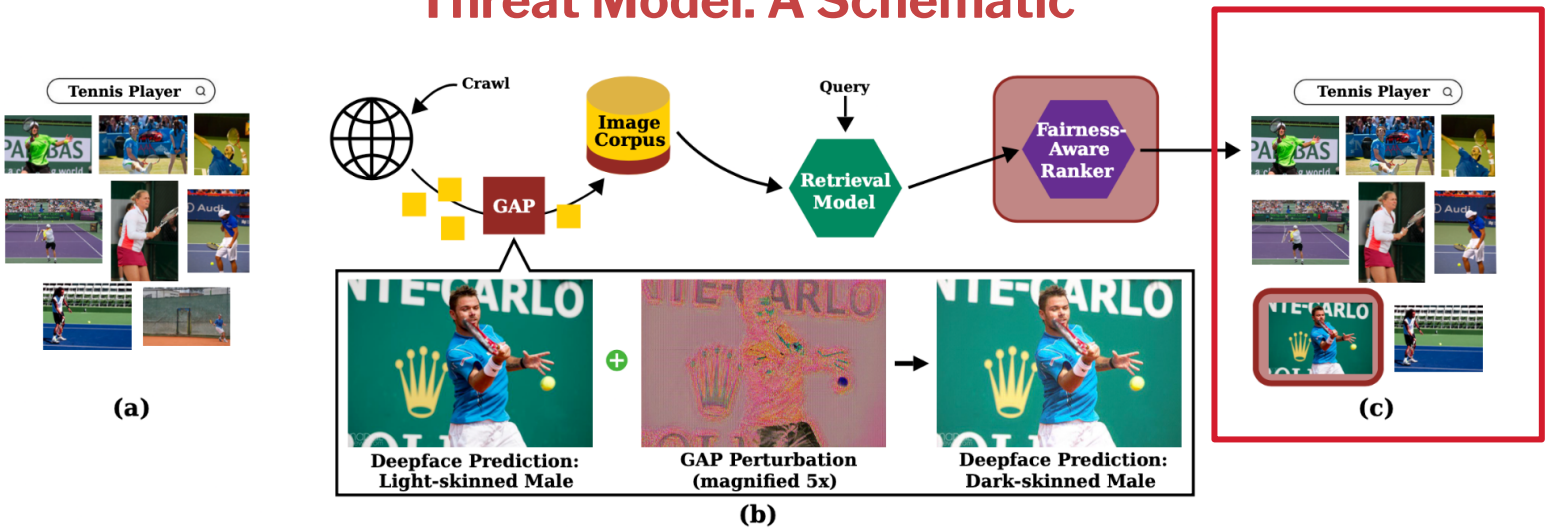
(a) shows example search results from an image search engine for the query “tennis player”.

Threat Model: A Schematic



(b) as this search engine crawls and indexes new images from the web, it collects images that have been adversarially perturbed using a GAP model

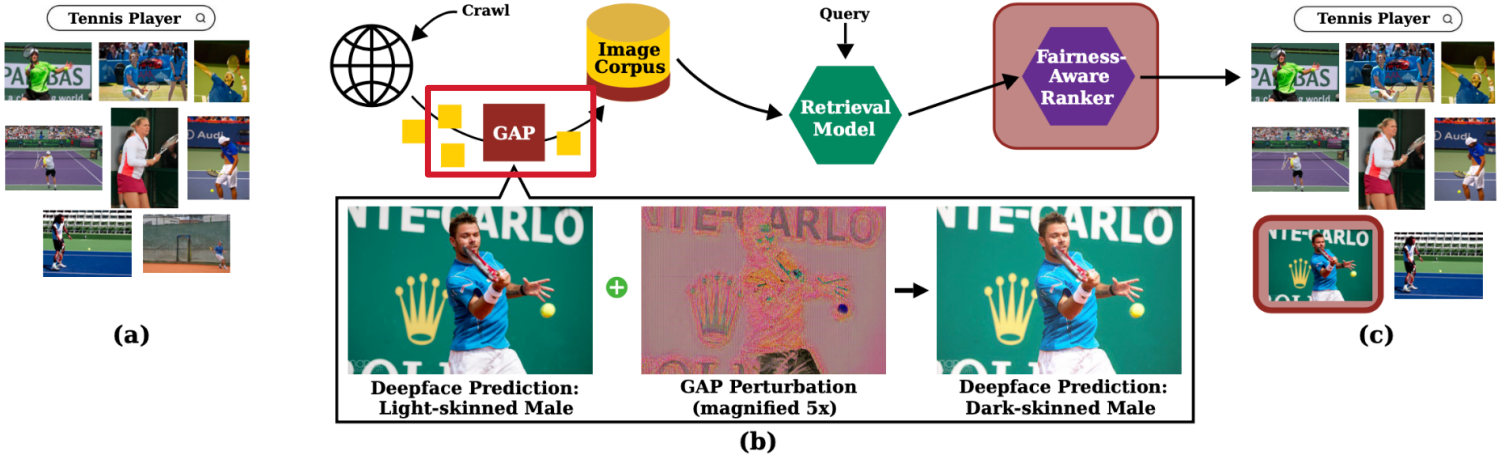
Threat Model: A Schematic



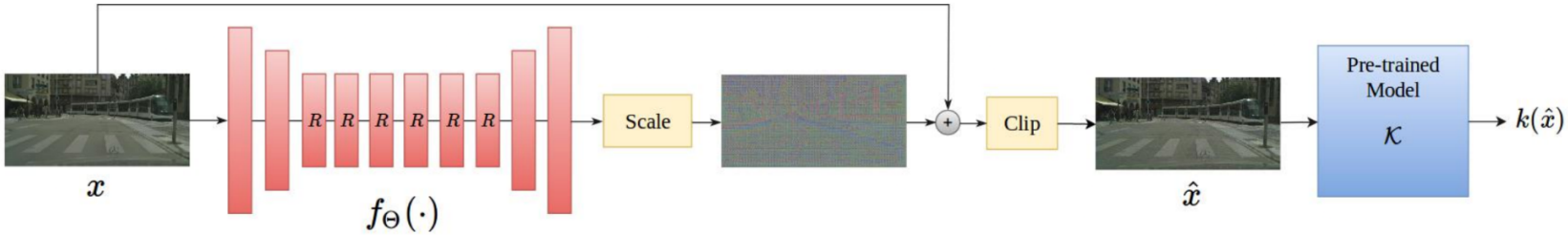
(c) The fairness-aware ranker (the target of the attack, highlighted in red) mistakenly elevates the rank of an image containing a light-skinned male (also highlighted in red) because it misclassifies them as dark-skinned due to the perturbations.

Methods

Setup: Generative Adversarial Perturbation

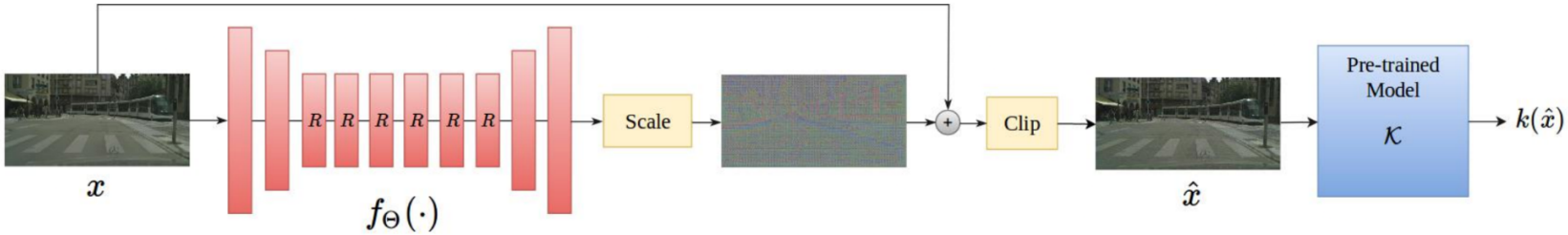


Setup: Generative Adversarial Perturbation



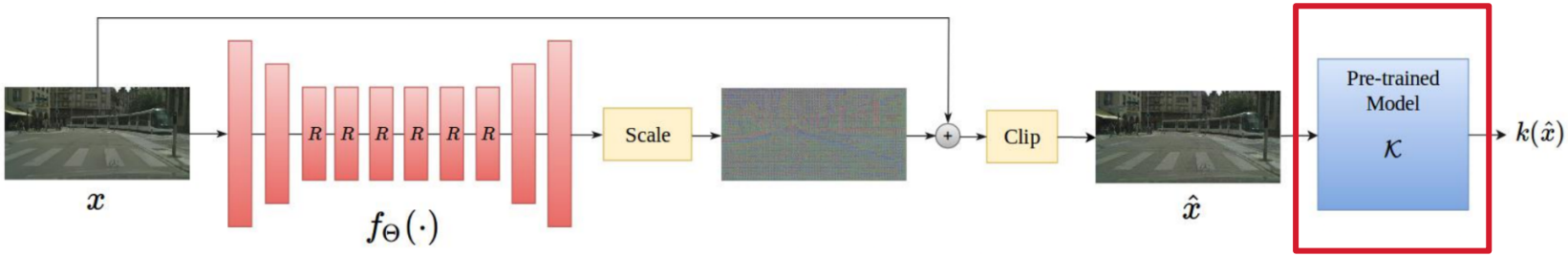
We modify a model called Generative Adversarial Perturbation (GAP) (Poursaeed et al. 2018). The adversary provides a source class y_s and target class y_t. The Class Targeted GAP model f_{CGAP} is a model that takes as input an image x and returns an image x′, effectively forcing the demographic inference model to misclassify samples of class y_s to class y_t, while maintaining its performance for samples not from class y_s .

Setup: Generative Adversarial Perturbation



A generative adversarial perturber model has advantages over universal perturbations because it does not require a fixed size or resolution image and can work on images of any size, which is what a realistic image search engine would be dealing with.

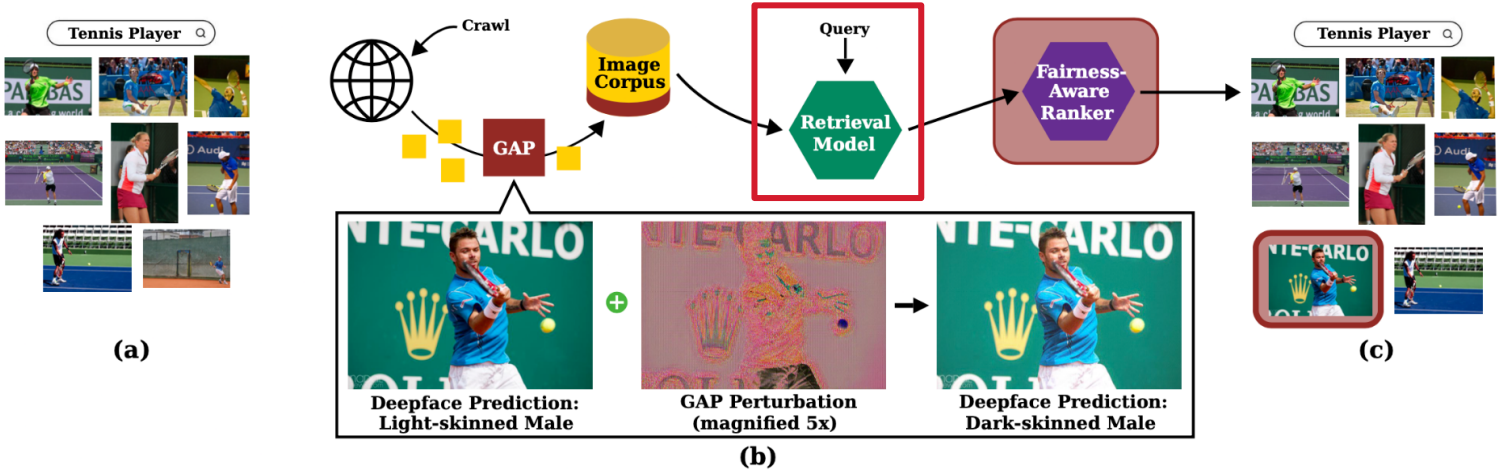
Setup: Generative Adversarial Perturbation



We use two pre-trained demographic classification models to train the GAP:

- **Deepface** is a face recognition model for gender and race inference developed by Facebook.
- **FairFace** is a model designed for race and gender inference, trained on a diverse set of 108,000 images.

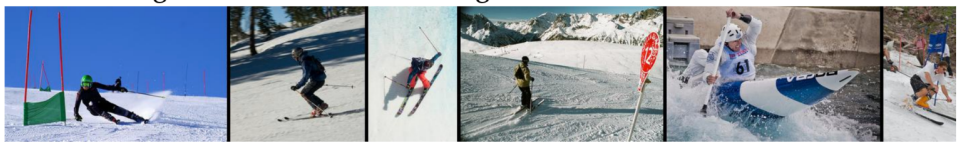
Setup: Retrieval Model



Setup: Retrieval Model

Caption: *A skier is skiing down the snow wearing a white shirt and black shorts.*

(a) Target Image

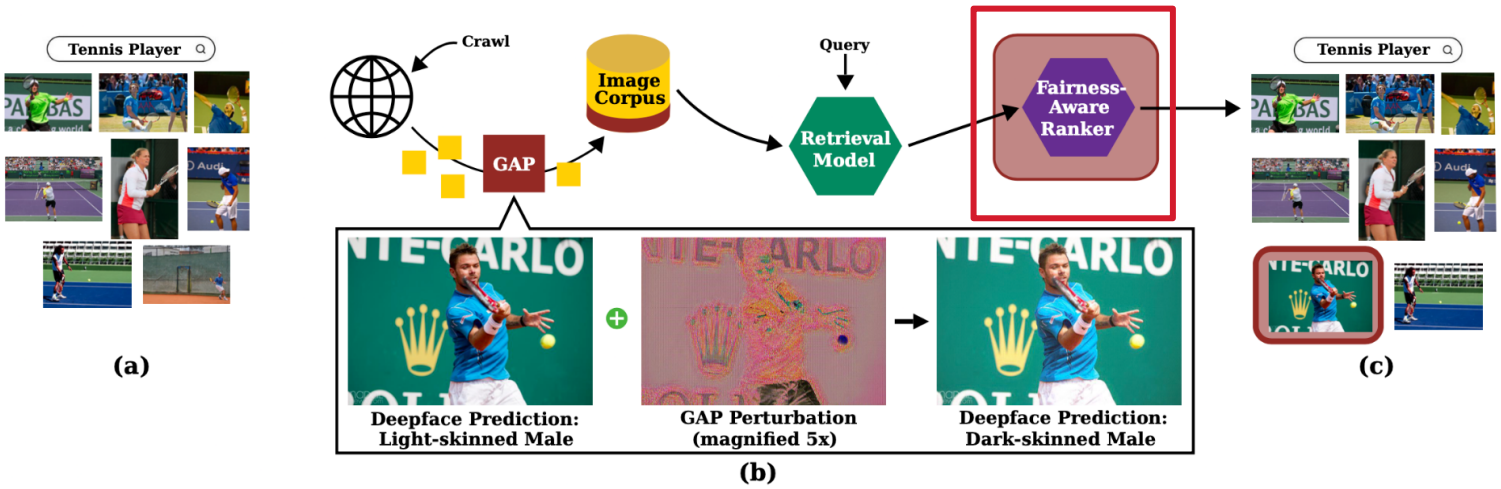


(b) Retrieved Top-6 with JOINT+BE^{OSCAR}

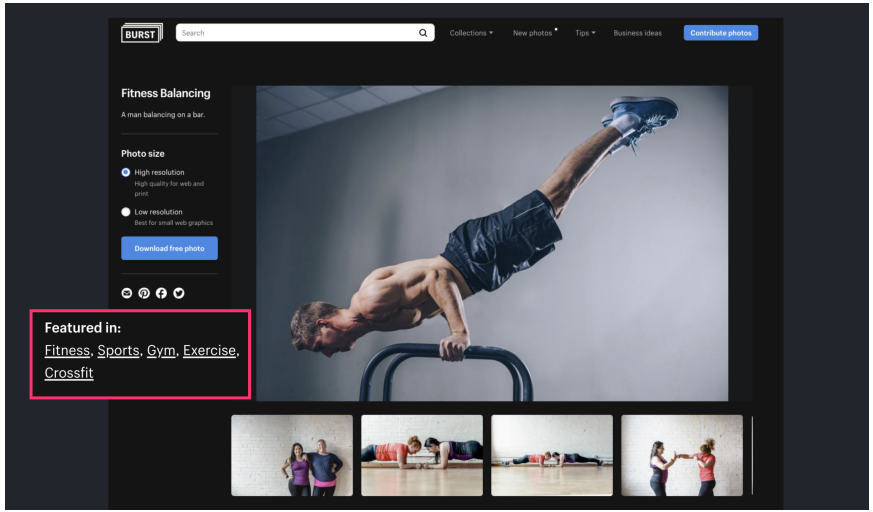
(c) Reranked Top-6 with JOINT+CE^{OSCAR}

The image search model we use in the paper is a MultiModal Transformer (MMT) (Geigle et al. 2021) based text-image retrieval model. This model consists of two components: a fast (although somewhat lower quality) retrieval step that identifies a large set of relevant images, followed by a re-ranking step that selects the best images from the retrieved set.

Setup: Fairness Aware Ranker



Setup: Fairness Aware Ranker



Two fair reranking models are evaluated in the paper:

- The LinkedIn **DetConstSort** algorithm, that uses protected attribute labels.
- Shopify's **Fair Maximal Marginal Relevance** (FMMR) algorithm, that encourages diversity in image embeddings.

Experiments

Setup: Case Study

Dataset: Skin color and gender annotated subset of **Microsoft COCO** (Zhao et Al.)

We only used images with one person. This amounted to 8692 images.

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Setup: Case Study

Search Queries	Attack Training	Training Objective	Attack Probability	Top k	Fair Ranking
“Tennis Player” “Person eating pizza” “Person at table”	Deepface FairFace	Any→Light Men Light Men→Any Dark Men→Light Men Light Men→Dark Men	0.2,0.5,0.7,1.0	10,15,20...,45,50	DetConstSort (LinkedIn) FMMR (Shopify)

Metrics

- **Skew** (Representation Bias), **Attention** (Position Bias), **NDCG** (Ranking Quality)
- We wanted to focus on the boost conferred to the majority subgroup - light men
- Summarizing metric: **Attack Effectiveness**

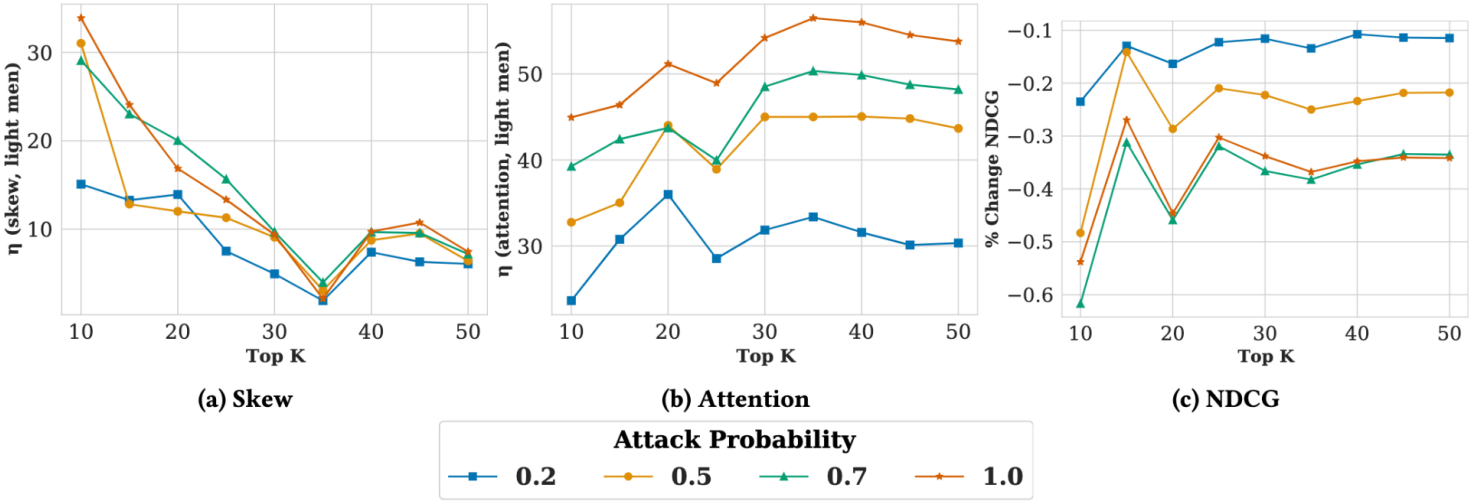
$$\eta(m, g) = \%\text{ change in } m \text{ for subgroup } g - \text{minimum \% change in } m \text{ over other subgroups.}$$

So, for example, η(attention, light men) measures the attack effectiveness by measuring the relative attention boost provided to light men after the attack.

Results

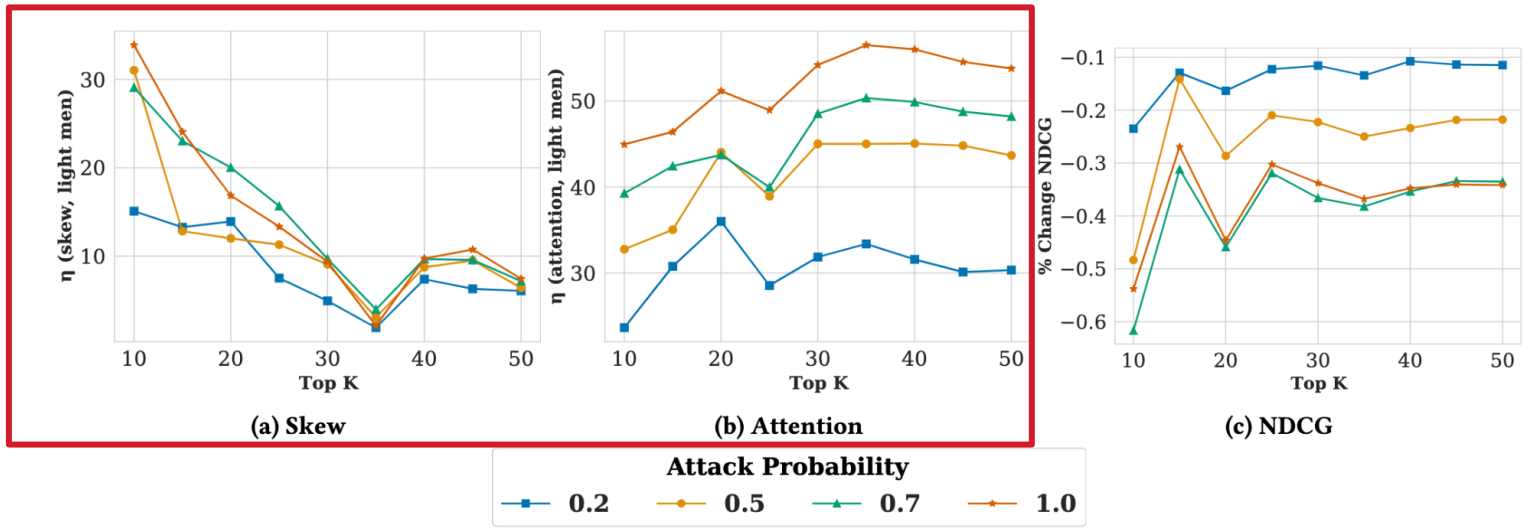
Results

Effect of Top K and Attack Probability



Results

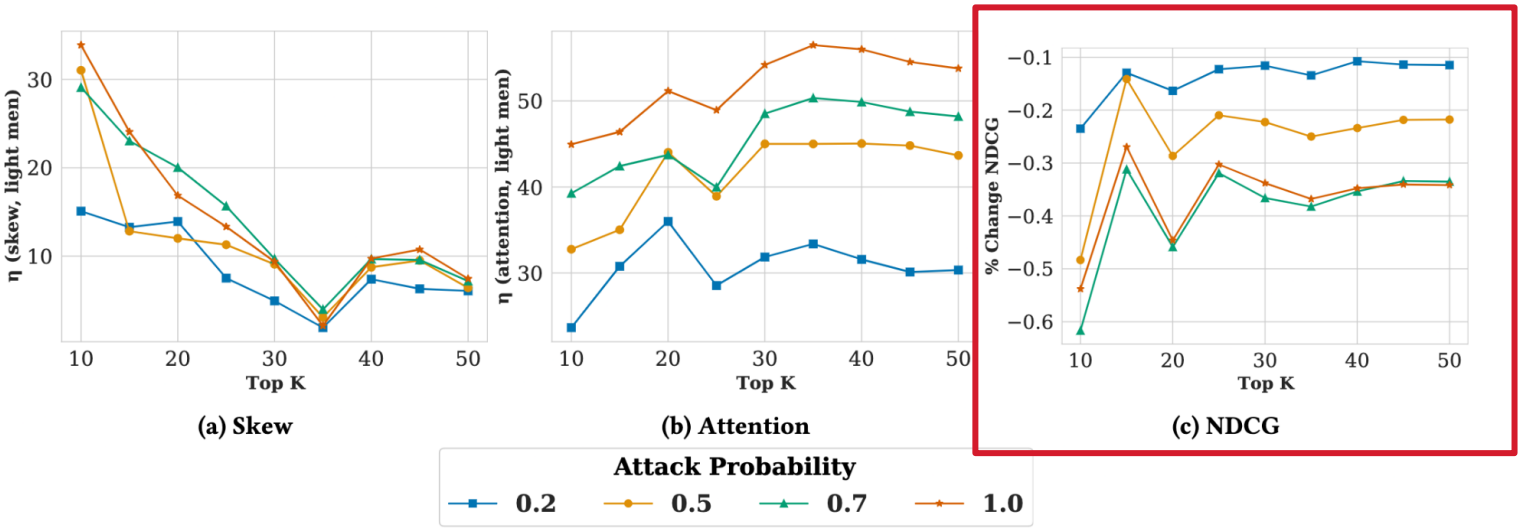
Effect of Top K and Attack Probability



Skew and Attention generally unfairly increase towards light men with increasing attack probability

Results

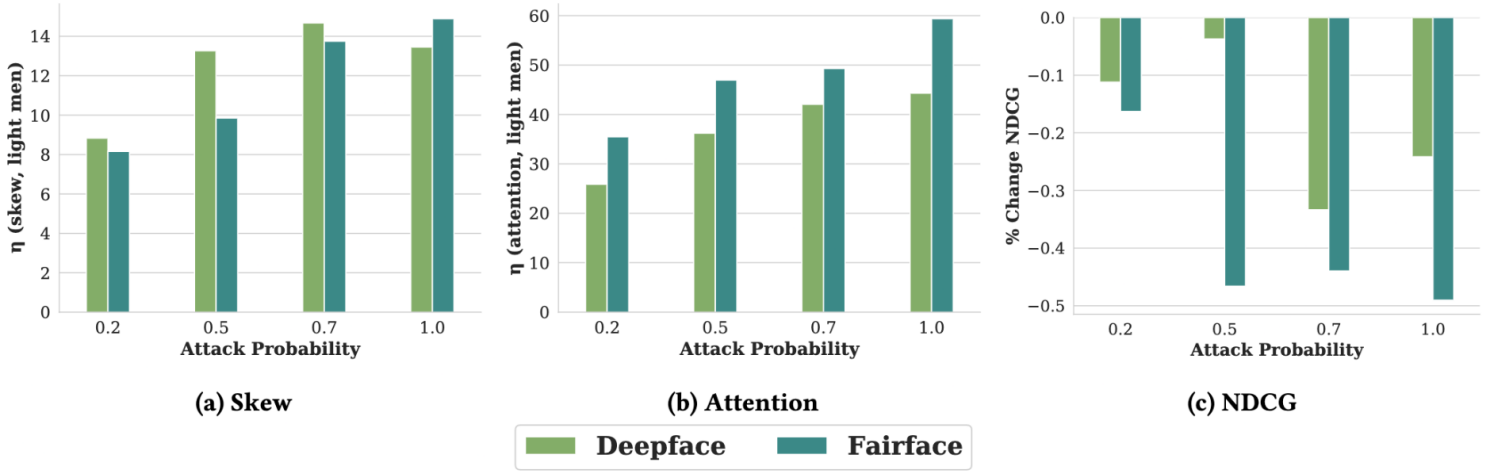
Effect of Top K and Attack Probability



Skew and Attention generally unfairly increases towards light men with increasing attack probability, however, NDCG is barely affected.

Results

Effect of Training Model



We observe that the attack effectiveness is similar, no matter what the model used for training.

Conclusion

- The attacks can successfully confer **significant unfair advantage to people from the majority class** (light-skinned men, in the case study)—in terms of their overall representation and position in search results—relative to fairly-ranked baseline search results.

- The attack is **robust** across a number of variables, including the length of search result lists, the fraction of images that the adversary is able to perturb, the fairness algorithm used by the search engine, the demographic inference algorithm used to train the GAP models, and the training objective of the GAP models.

- The attacks are **stealthy**, i.e., they have close to zero impact on the relevance of search results, and the perturbations are invisible to the human eye.


That's all Folks!

Thank you.
Questions?

Co-authors:

Matthew Jagielski
Google Brain

Christo Wilson
Northeastern University