SoK: Anti-Facial Recognition Technology

Emily Wenger, Shawn Shan, Haitao Zheng, Ben Y. Zhao
Department of Computer Science, The University of Chicago
{ewenger, shansixioing, htzheng, ravenben}@uchicago.edu

Abstract—**The rapid adoption of facial recognition (FR) technology by both government and commercial entities in recent years has raised concerns about civil liberties and privacy. In response, a broad suite of so-called "anti-facial recognition" (AFR) tools has been developed to help users avoid unwanted facial recognition. The set of AFR tools proposed in the last few years is wide-ranging and rapidly evolving, necessitating a step back to consider the broader design space of AFR systems and long-term challenges. This paper aims to fill that gap and provides the first comprehensive analysis of the AFR research landscape. Using the operational stages of FR systems as a starting point, we create a systematic framework for analyzing the benefits and tradeoffs of different AFR approaches. We then consider both technical and social challenges facing AFR tools and propose directions for future research in this field.**

I. Introduction

In recent years, facial recognition systems have accelerated their growth in scale and reach, becoming an increasingly ubiquitous part of our daily lives. The majority of citizens in the world's most populous countries are enrolled in one or more facial recognition systems, whether they know it or not. In the United States, nearly 200 million Americans are enrolled in the FBI facial recognition database, which leverages access to driver license photos from most states [1]. In China, a well-known surveillance system uses facial recognition to monitor civilian behavior and enforce the social credit score system [2], [3]. In Russia, authorities acquired 100,000+ cameras in Moscow to build a facial recognition-based COVID quarantine enforcement system [4]. Beyond government use cases, facial recognition systems are now regularly used for myriad purposes, including authenticating travelers at airports and employees entering corporate offices.

The advancements that paved the way to these facial recognition systems have also opened the door to their potential misuse and abuse. With moderate resources, an individual or institution, public or private, can now extract training data from social media and online sources to build facial recognition models capable of recognizing large groups of users. In 2020, New York Times journalist Kashmir Hill demonstrated the potential for facial recognition misuse when she profiled Clearview.AI, a private for-profit company that scraped over 3 billion images from "public sources" to build a facial recognition system that recognized hundreds of millions of private citizens [5], without their knowledge or consent. Clearview and companies like it could enable surveillance and tracking by anyone willing to pay[1]. In addition to images shared online, other reports have detailed how photos taken in unexpected places – airports, city streets, government buildings, schools, corporate offices – can end up in facial recognition systems without subjects' knowledge or consent (e.g., [1], [7], [8], [9], [10], [11]).

Despite backlash against intrusive facial recognition systems [12], [13], [14], [15], there are few tools available to protect users against them. While big tech has begun to self-regulate [16] and openly called for legislation [13], [12], legislative efforts to regulate facial recognition remain scarce. In their place, a cottage industry of anti-facial recognition (AFR) tools has emerged. These AFR tools are designed to target different parts of facial recognition systems, from data collection and model training to inference, with the unified goal of preventing successful recognition by unwanted or unauthorized models.

In the last 12 months, more than a dozen AFR tools have been proposed (e.g., [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32]). While most are constrained to research prototypes, a few of these tools have produced public software releases and gained significant media attention [19], [22], [33].

Proposals in the rapidly growing collection of AFR tools differ widely in their assumptions and techniques and target different pieces of the facial recognition pipeline. There is a need to better understand their commonalities, to highlight performance tradeoffs, and to identify unexplored areas for future development. In this paper, we address this need, through the lens of a common framework for analyzing a wide range of AFR systems.

More specifically, we make the following contributions:

- **Taxonomization of targets in facial recognition:** AFR systems target a wide range of components in the facial recognition process. Using a generalized version of the facial recognition data pipeline, we provide the first framework to reason broadly about existing and future work in this space.
- **Categorization and analysis of AFR systems:** We take the current body of work on AFR systems, categorize and analyze them using our proposed framework.
- **Mapping design space based on desired properties:** We identify a core set of key properties that future AFR systems might optimize for in their design, and provide a design roadmap by discussing how and if such properties can be achieved by AFR systems that target each stage in our design framework.
- **Open challenges:** We use our framework to identify sig-

[1] Multiple countries are pursuing inquiries into Clearview's business model, and Canada has already denounced it as "mass surveillance" and "illegal" [6].

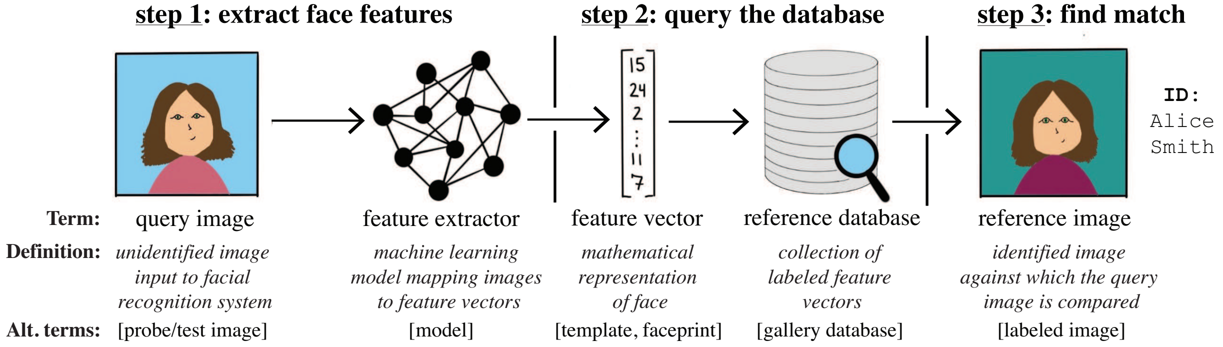
step 1: extract face features
step 2: query the database
step 3: find match
15
24
2
:
11
7
ID:
Alice
Smith
Term:
query image
feature extractor
feature vector
reference database
reference image
Definition:
unidentified image input to facial recognition system
machine learning model mapping images to feature vectors
mathematical representation of face
collection of labeled feature vectors
identified image against which the query image is compared
Alt. terms:
[probe/test image]
[model]
[template, faceprint]
[gallery database]
[labeled image]

Fig. 1. The workflow of how facial recognition systems recognize a human face in an input image, along with the corresponding terminology. (a): A query image, after being submitted to the system, is passed to the feature extractor to produce a feature vector; (b): this feature vector is used to query a reference database of labeled feature vectors; (c): if the query feature vector matches a labeled feature vector in the database, the label is used to find a reference image, and the system outputs the reference image and the identity (i.e. Alice Smith in this example).

nificant challenges facing current AFR systems, as well as directions for potential solutions.

In the rest of the paper, we begin with a detailed description of real-world facial recognition systems (§II), including real-world deployment scenarios and key technical components. We then present the motivation and threat model of AFR systems (§III), and our systemization of existing AFR tools by examining the five overarching stages of facial recognition systems that AFR tools could target (§IV). We discuss the key attack methods used by existing AFR proposals targeting each stage, i.e., *data collection* (§V), *data processing* (§VI), *feature extractor training* (§VII), *identity creation* (§VIII), and *query matching* (§IX). We then identify key desirable properties for future AFR systems, and map them to points in the design space (§X). Finally, we discuss open challenges and potential directions for future AFR research (§XI).

Unresolved Ethical Questions: The broad deployment of facial recognition systems (and by extension, AFR systems) is fraught with ethical challenges and implications, not the least of which are significant biases against women and people of color [34]. While we discuss ethical tensions surrounding AFR systems in §XI-B, we do not make assertions in this paper on how (and whether) AFR tools should be used. Development and adoption of AFR tools are driven by backlash against biases in and misuse of facial recognition systems. Even as we continue to struggle with their legal and ethical implications, we recognize that AFR tools are here to stay, and an analysis of their strengths and limitations is crucial to advancing the ongoing debate about both their use and the place of facial recognition in our world.

II. Facial Recognition: Workflow, Design Stages and Deployment

As context for later discussions, we now provide an overview of facial recognition (FR) systems and their real-world implementations. FR systems identify people by their facial characteristics, generally by comparing an unidentified human face in an image or a video against a database of facial images with known identities. While there are many design variants [35], we focus on the state-of-the-art and widely adopted FR systems, which employ deep neural networks (DNNs) to perform recognition on digital face images.

We note the distinction between *facial recognition* systems, the main target of AFR systems and the subject domain of this work, versus *facial verification* systems. Facial verification is used widely to authenticate users on mobile devices (e.g. FaceID on iPhones), by checking the similarity of a user's facial features against the stored feature vector matching the authorized user. The large majority of AFR systems focus only on facial recognition, and as such, we do not consider facial verification or its disruption in this work.

Below, we begin by presenting the run-time workflow of facial recognition. We then propose a breakdown of the FR workflow into five *operational stages*, a framework that we will revisit and use for analyzing AFR systems in Section IV. Finally, we give an overview of real-world deployments of FR.

A. Run-time Facial Recognition Workflow

Figure 1 summarizes the run-time workflow of how FR systems identify a face from an input image. *First*, a *query image*, i.e. a face image to be identified, is fed through a *feature extractor*, a DNN that converts the image into a *feature vector* (or a mathematical representation of the person's facial features). *Next*, this feature vector is used to query a *reference database*, a collection of face images of known identities. This query search is done by comparing the input feature vector against the reference feature vectors stored in the database to find the closest match. *Finally*, if the query search finds a reference feature vector in the database sufficiently similar to the input, the FR system declares that a match has been found and outputs the corresponding identity and the associated *reference image* (i.e. Alice Smith in Figure 1).

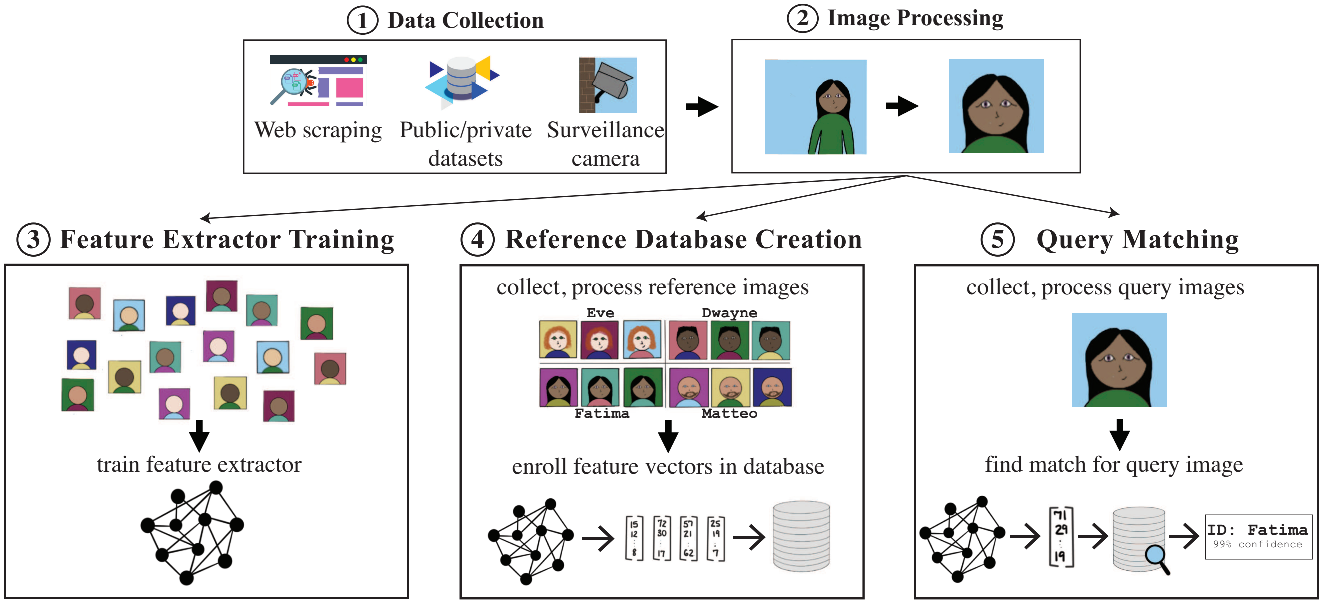
① Data Collection
Web scraping
Public/private datasets
Surveillance camera
② Image Processing
③ Feature Extractor Training
train feature extractor
④ Reference Database Creation
collect, process reference images
Eve
Dwayne
Fatima
Matteo
enroll feature vectors in database
⑤ Query Matching
collect, process query images
find match for query image
ID: Fatima
99% confidence

Fig. 2. We propose to divide the facial recognition operational pipeline into a set of five operational stages (① to ⑤). They encompass the five critical points of direction interaction between users and FR systems. Later we will use this framework for analyzing AFR systems.

It is worth noting that the terminology used to describe a FR system can vary across the literature, and some alternative terms are listed in Figure 1. For example, *query images* are sometimes called "probe images" or "test images," while *feature vectors* are referred to elsewhere as "face templates" or "faceprints". *Reference images* are also known as "identified images" or "gallery images". The terms we choose to use in this paper are, we believe, most familiar to the security research community.

B. Breaking FR into Operational Stages

We now examine the FR operational pipeline and divide it into a set of *operational stages* that will frame our discussion of FR and AFR tools. These operational stages correspond to specific subtasks in FR, which together encompass the five critical points of direct interaction between users and FR systems. Figure 2 depicts the five operational stages of FR. We discuss each stage below, and will revisit them as a framework to analyze anti-facial recognition tools in §IV.

① **Image collection.** Face images primarily come from two sources: online image scraping [61] or physically taking a photo of a person [1], [8]. We discuss sources of face images for FR systems in further detail in §II-C.

② **Image preprocessing.** Raw images from stage ① are often poorly structured (e.g., varying face sizes, bystanders in background). To make downstream tasks easier, the FR system often preprocess images by applying face detection (e.g., automated face cropper [62]) to remove the background and extract each individual face, followed by a data normalization process [63], [64], [65].

③ **Training feature extractor.** The crucial element of DNN-based FR systems is the feature extractor used to compute facial features from an image. To achieve accurate recognition, the computed feature vectors must be highly similar for photos of the same person, but sufficiently dissimilar across photos of different people. To enable this behavior, most existing FR systems adopt the training methodology proposed by [65] in 2015: adding an *extra* loss function during model training to directly optimize for large separations between different faces in the feature space. Followup works explore alternative loss functions and model architectures to further improve the accuracy of FR systems (e.g., [63], [64], [66]).

To maximize efficacy, the feature extractor is generally trained on millions of labeled face images. Extensive resources are required to both collect and label a large face dataset and to actually train the model. As a result, many FR practitioners, including large companies [67] and government agencies [68], [69], opt to purchase or license a well-trained feature extractor from tech companies (e.g. [70], [71], [72], [73], [74], [75], [76], [77]).

④ **Reference database creation.** FR systems need a large database of known (or labeled) faces in order to match unknown (unlabeled) faces to their true identities. As a result, FR systems build a reference database of people they want to recognize, by first collecting and preprocessing labeled face images of these individuals, and then passing them to the feature extractor to obtain feature vectors. The reference database stores the corresponding feature vector and identity pairs [78], [61], [79].

⑤ **Query matching.** At run-time, the FR system takes in an unidentified face image, extracts its feature vector, then

Type	Use Cases Reported	Countries/Companies
Government	On-street surveillance	Bahrain [36], China [2], England [15], France [37], Kenya [38], Myanmar [39], Russia [4], UAE [36], UK [40], US [9], Zimbabwe [41]
	Criminal suspect identification	Argentina [42], Belarus [39], Brazil [43], China [44], Greece [39], Malaysia [45], US [1]
	School monitoring	Brazil [46], China [47], India [11], Russia [46], US [46]
	Border security	Israel [48], Pakistan [49], US [50]
	COVID lockdown enforcement	China [51], India [52], South Korea [52], Russia [4]
Commercial	Catching shoplifters	Apple, Macy's, Lowe's [53], [54]
	Securing facility access	Alibaba [55], Intel [56]
	Tracking driver behavior	Hyundai [57], Subaru [58]
	Air passenger check-in	JetBlue [59], Delta [60]

TABLE I
EXAMPLE USE CASES OF FACIAL RECOGNITION.

uses it to query the reference database to locate a match (if any exists). If the feature space distance (e.g., L_2 or cosine) of the query image is close enough to an entity in the database, the system outputs a match.

C. Real World FR Deployment and Data Collection

In recent years, large corporations and government agencies across the globe have adopted FR for various applications. This wide adoption was triggered by significant accuracy improvements of FR systems, largely due to new training methods [65] and more powerful neural network architectures [80]. Below, we present some commonly known FR use cases and discuss their impact on users.

Government use cases. Government agencies around the globe use FR for a variety of purposes. For example, the US government uses FR systems for law enforcement purposes such as border control [50] and police operation[2] [81]. The Chinese government employs FR to monitor specific subpopulations [2], [82], track video game use [83], and enforce COVID lockdowns [51]. Table I lists more examples of government uses of FR. For a broader exploration of this topic, we refer the reader to [84].

Commercial use cases. Many corporations have integrated FR into their security and commerce pipelines. The most common FR use cases are enhancing store or office security. For example, companies like Apple, Macy's, and Lowes have begun using FR to catch shoplifters in their stores [54]. Other companies have employed FR to monitor corporate facility access [55], [56]. Product-based applications have emerged as well, such as car companies like Subaru using FR to track driver fatigue [58] or airlines using FR to streamline passenger checkins [59], [60].

Sources of face images. The definitive source of images for deployed FR models is often unknown. Based on government reports and media articles, we outline some known sources of training, reference, and query images used by today's FR systems.

Training images (used to train feature extractors) often come from a mix of academic training datasets (*e.g.* [85], [86], [87], [88]), proprietary data, and public data scraped from social media accounts, according to a report of the US Government Accountability Office [68]. Reference images used to create the reference database generally come from the Internet (e.g., social media), or government databases (e.g., passport and driver license photos). A list of known reference image sources for some well-known FR operators is shown in Table II. Finally, query images can come from both online and physical sources, including social media, police body cams, mug shots, corporate surveillance systems, state identification images, passport photos, and others [69].

Operator of FR system	Source of reference images
Clearview.ai	Social media photos [90]
PimEyes	(Public) online photos [91]
FBI F.A.C.E.S.	State drivers' license photos [1]
US Customs and Border Patrol	Passport photos [50]
Skynet (China)	National ID photos [92], [3]

TABLE II
REPORTED REFERENCE IMAGE SOURCES

After identification, query images are often fed back into the reference database, either to enhance existing feature vectors or create new ones. For example, US Customs and Border Patrol states that images of non-US travelers collected at US entry points are fed back into a larger DHS database as reference images. Similar techniques are used by several Chinese companies [89], [10].

III. Anti-Facial Recognition: Motivation and Threat Model

In this section, we discuss factors driving the development of anti-facial recognition (AFR) tools, the threat model of those AFR tools, and its practical implications.

A. The Rise of AFR Tools

Numerous forces have coalesced to drive the recent trend in AFR tool development. *First*, numerous reports about the provenance of images used in commercial FR systems have raised significant privacy concerns. The most infamous examples are Clearview.ai and PimEyes – both companies have scraped over *3 billion images* from social media sites to use in their FR systems [90] without user knowledge or consent. *Second*, increased government use of FR systems has caught the eye of citizens who have raised significant concerns about

[2] Recently, police departments around the US have drawn fire for their use of highly unregulated FR software like Clearview.ai [71].

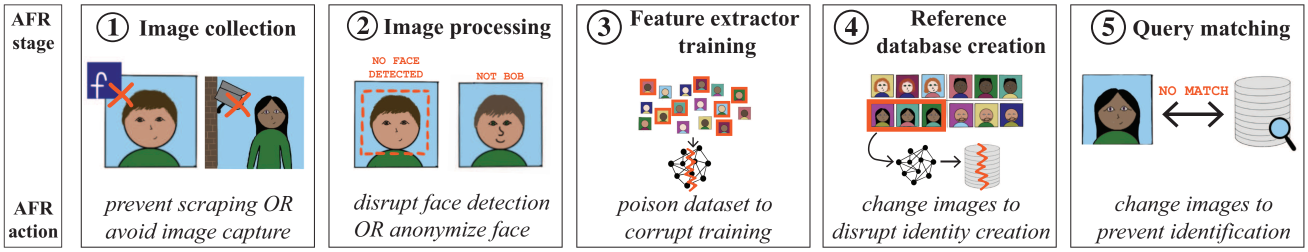
AFR stage
AFR action
① Image collection
prevent scraping OR avoid image capture
② Image processing
NO FACE DETECTED
NOT BOB
disrupt face detection OR anonymize face
③ Feature extractor training
poison dataset to corrupt training
④ Reference database creation
change images to disrupt identity creation
⑤ Query matching
NO MATCH
change images to prevent identification

Fig. 3. Overview of our proposed stage-based framework for analyzing existing AFR proposals. We list the five critical stages of facial recognition as discussed in §II-B and present AFR strategies per stage by the attack target, action, and desired effect.

the long-term effects of FR on privacy and freedom of expression [15], [93]. *Third*, multiple editorials have highlighted and discussed the demographic bias of existing FR systems, calling for a moratorium on (or at least regulation of) the FR technology [13], [94], [95].

Consequently, public sentiment about FR is mixed and, especially in western countries, trending negative [96], [97], [98], [99]. This shift in public opinion, combined with the concerns and forces noted above, has motivated researchers to create various AFR tools to counteract unwanted FR systems.

B. Threat Model of AFR

AFR tools are used by a person $\mathcal{P}$ to combat a FR system or service $\mathcal{F}$. In this context, $\mathcal{P}$ takes the role of an attacker and acts against $\mathcal{F}$. Development of AFR tools generally makes the following assumptions about each party:

- $\mathcal{P}$ has no special access to or authority over the target FR system $\mathcal{F}$, but wishes to evade unwanted facial recognition by modifying or otherwise controlling their own face images.
- $\mathcal{F}$'s goal is to either create or maintain an accurate facial recognition operation. Furthermore, $\mathcal{F}$ operates *at scale* and does not specifically target $\mathcal{P}$ for identification.

Assumptions and Implications. The above threat model relies on several key assumptions. We now discuss their implications.

1) *Assuming AFR tools operate on images*: Our study focuses exclusively on image-based AFR tools that a user $\mathcal{P}$ can deploy themself. These image-based designs dominate the current set of AFR proposals. Yet a user $\mathcal{P}$ may, depending on their context, be able to use other means (*e.g.*, legal action) to fight unwanted facial recognition. We discuss potential non-image-based AFR methods later in §XI.

2) *Assuming $\mathcal{F}$ does not specifically target $\mathcal{P}$ for recognition*: We note that existing AFR tools are designed to fight large-scale FR systems. This is because, from a practical standpoint, if system $\mathcal{F}$ wishes to specifically recognize a user $\mathcal{P}$, there are much more efficient options than using a general, large-scale FR system. Therefore, most AFR tools are not designed to withstand this level of scrutiny.

User Control	Data Source
High	Photos taken in academic research study. Signed release for photos taken at public event.
Medium	Photos posted online by user on personal social media.
Low	Photos posted online by user's friends. Images sold by companies without user knowledge. Photos obtained from surveillance cameras in public spaces. Photos from government databases.

TABLE III
A LIST OF COMMON SCENARIOS WHERE A USER'S LEVEL OF CONTROL OVER THEIR FACE IMAGES VARIES.

Section	Description
§III	Motivation and threat model of AFR tools
§IV	Overview of AFR strategies + taxonomy
§V–§IX	Details of AFR proposals targeting each FR stage
§X	Benefits and limitations of attacking each FR stage
§XI	Challenges facing AFR development Potential future directions

TABLE IV
OVERVIEW OF OUR SYSTEMIZATION AND ANALYSIS OF AFR PROPOSALS

IV. A Stage-based Framework for Analyzing AFR

We now discuss and analyze existing AFR proposals. To do so, we propose and use a *stage-based* framework to categorize AFR strategies, which encompasses the five critical points of direct interactions between users and FR systems. AFR tools can operate at these points, where FR systems interface with the broader world.

As shown in Figure 3, each of these critical points corresponds to an key operational stage of FR systems, i.e. the stages ① – ⑤ described in §II. With this in mind, we now summarize the "attack" strategies used by AFR tools to disrupt the operation of each FR stage and taxonomize current AFR proposals. In the next few sections, we discuss in detail the AFR proposals targeting each individual stage (§V–§IX), before discussing the goals and tradeoffs of AFR tools (§X). Finally, we consider broad challenges facing AFR development and discuss potential future directions(§XI). The overall structure of our analysis is shown in Table IV.

A. AFR Strategies per Stage

Since the five FR stages ①–⑤ encompass the points of direct interaction between $\mathcal{P}$ and $\mathcal{F}$, they naturally cover the points of attack employed by existing AFR proposals. Next we briefly describe the general strategies used by AFR tools targeting each FR stage.

AFR system	Year released	Stage targeted	Attack scenario				
			$\mathcal{P}$'s knowledge of $\mathcal{F}$	*$\mathcal{P}$'s operating context*	*Targeted/ Untargeted*	*Tested on real-world FR*	Unique Property
Anti-scraping [115-119]	2021	①	-	Digital	UT	-	Prevent large-scale image scraping
Data Leverage [100]	2021	①	-	Digital	UT	-	Withholds data to prevent collection.
CVDazzle [33]	2010	2a	WB	Physical	UT	-	Make-up
Xu et al. [26]	2020	2a	BB	Physical	UT	YOLOv2	Adversarial patch on T-shirts
Wu et al. [24]	2020	2a	Both	Physical	UT	YOLOv2	Adversarial patch on T-shirts
Zolfi et al. [101]	2020	2a	BB	Physical	UT	YOLOv5	Stickers on camera lens that blur vision
SocialGuard [28]	2020	2a	WB	Digital	UT	-	Adversarial perturbation on face detectors
Treu et al. [25]	2021	2a	BB	Digital	UT	-	Adversarial clothing on face detectors
DeepPrivacy [102]	2019	2b	BB	Digital	UT	-	GAN-based face blurring (perceptible)
IdentityDP [18]	2021	2b	BB	Digital	UT	AZ	GAN-based face blurring (perceptible)
DeepBlur [17]	2021	2b	BB	Digital	UT	AZ, F++	GAN-based face blurring (perceptible)
Yang et al [103]	2021	2b	BB	Digital	UT	-	GAN-based face blurring (imperceptible)
Evtimov et al. [104]	2021	③	BB	Digital	UT	-	Data poison by modifying entire dataset
Huang et al. [20]	2021	③	BB	Digital	UT	-	Data poison by user coordination
Fawkes [19]	2020	④	Both	Digital	UT	AR, AZ, F++	Corrupts features of faces
FoggySight [21]	2021	④	Both	Digital	UT	AZ	Collectively corrupts features of faces
LowKey [22]	2021	④	BB	Digital	UT	AR, AZ	Corrupts features of faces
Feng et al. [105]	2013	⑤	BB	Physical	UT	-	Make-up
Sharif et al. [106]	2016	⑤	Both	Both	Both	F++	Adversarial patch on wearable accessories
Dabouei et al. [107]	2018	⑤	WB	Digital	UT	-	Adversarial attack distorts face landmarks.
Zhou et al. [108]	2018	⑤	WB	Physical	Both	-	Projected adversarial IR patterns
Dong et al. [109]	2019	⑤	BB	Digital	T	TN	Black-box adversarial perturbation.
Zhu et al. [110]	2019	⑤	Both	Digital	Both	-	Adds eye makeup with GAN.
AdvHat [30]	2019	⑤	WB	Physical	UT	-	Printed sticker on hat.
AdvFaces [111]	2019	⑤	BB	Digital	Both	-	GAN-based adversarial attack.
VLA [112]	2019	⑤	BB	Physical	Both	-	Projected light patterns
Nguyen et al. [29]	2020	⑤	Both	Physical	Both	?	Projected light patterns
Browne et al. [31]	2020	⑤	BB	Digital	UT	-	Universal adversarial perturbation
Cilloni et al. [23]	2020	⑤	WB	Digital	UT	-	Corrupts features of faces
Face-Off [27]	2020	⑤	BB	Digital	Both	AR, AZ, F++	Study on user perception on perturbation levels.
Singh et al. [113]	2021	⑤	WB	Digital	UT	-	Brightness-agnostic adversarial perturbations
Yang et al [114]	2021	⑤	BB	Digital	UT	TN	Corrupts features of faces

TABLE V

TAXONOMY OF PROPOSED AFR TOOLS. "BB/WB" = BLACK BOX, WHITE BOX."UT, T" = UNTARGETED, TARGETED. "AR, AZ, F++, TN" = AMAZON REKOGNITION, MICROSOFT AZURE FACE RECOGNITION, MEGVII'S FACE++, TENCENT FACE RECOGNITION.

Attacking ①. In the image collection stage, labeled and/or unlabeled images are collected for use by $\mathcal{F}$, either by physically taking photos or scraping online images. When targeting this stage, AFR tools focus on disrupting the data collection process to prevent $\mathcal{F}$ from acquiring usable face images of $\mathcal{P}$.

Attacking ②. This second stage pre-processes collected face images using a series of digital transformations, e.g., face detection, background cropping, and normalization. AFR tools deployed at this stage seek to render the processed images unusable, either by breaking the preprocessing functions (e.g., preventing faces from being detected), injecting noise and artifacts onto the images, or removing $\mathcal{P}$'s identity information from the images.

Attacking ③. Since stage ③ is dedicated to training face feature extractors, AFR tools targeting this stage seek to degrade the accuracy of the extractor by poisoning its training images.

Attacking ④. To create the reference database, labeled reference images are passed through the feature extractor to create their feature vectors. AFR tools targeting this stage attempt to corrupt the feature vectors created for $\mathcal{P}$'s reference images so that the database holds a "wrong" feature vector of $\mathcal{P}$.

Attacking ⑤. In the query matching stage, AFR tools seek to prevent accurate matching between a query image's feature vector (of $\mathcal{P}$) and $\mathcal{P}$'s feature vectors stored in $\mathcal{F}$'s reference database. This is generally achieved by perturbing (or modifying) the query image to change its feature vector.

B. Taxonomy of Existing AFR Proposals

Using our stage-based analysis framework, we now present a comprehensive taxonomy of existing AFR proposals in Table V. In this list, we categorize existing AFR proposals by the year of release, the individual FR stage they target, and the attack scenario. We further break down the attack scenario by $\mathcal{P}$'s knowledge of $\mathcal{F}$ (white box or black box[3]), the AFR deployment context (physical or digital), whether the attack is targeted or untargeted[4], whether the AFR tool has

[3] *White box* means $\mathcal{P}$ has full knowledge of and access to $\mathcal{F}$'s FR system (including feature extractor parameters) and uses this knowledge to guide their AFR protection. *Black box* means $\mathcal{P}$ lacks such access and knowledge.

[4] A *targeted* attack causes the FR system to identify $\mathcal{P}$ as a specific, incorrect person (e.g. a famous politician). An *untargeted* attack means that $\mathcal{P}$ is misclassified, but not as a specific person.

been tested against real-world FR systems, and any unique or notable features of the AFR tool.

We note a significant imbalance of AFR tools targeting different stages. Stage ② and ⑤ have attracted the most number of AFR proposals, likely due to the popularity of adversarial perturbation-based research. We also notice that 7 out of 30 proposals assume a "white-box" access to $\mathcal{F}$'s FR pipeline, which is often unrealistic in practice. Finally, only 12 out of the 30 proposals have tested the AFR effectiveness against at least one real-world FR system. Overall, Table V serves as a comprehensive summary of current AFR proposals, which we will refer to throughout the paper.

V. Attacking ① to Disrupt Data Collection

In the next five sections, we discuss in greater detail how existing AFR proposals attack each of the five stages. In each section, we first describe the goals of $\mathcal{F}$ and $\mathcal{P}$ in the corresponding stage and then discuss specific AFR proposals that allow $\mathcal{P}$ to disrupt $\mathcal{F}$.

This section focuses on methods that allow $\mathcal{P}$ to attack $\mathcal{F}$ by disrupting the process of face data collection (stage ①).

- **$\mathcal{F}$'s goal** is to obtain usable face images from online or physical sources. In many scenarios, $\mathcal{F}$ aims to collect high quality images of millions or billions of people (e.g., Clearview.ai [90]).
- **$\mathcal{P}$'s goal** is to prevent their face images from being collected for use in face recognition systems. They use online or physical evasion/disruption techniques to thwart image collection.

Face images can come from two sources: scraping online images or physically capturing faces using cameras. Thus we divide existing AFR tools acting at this stage into two subcategories: preventing scraping of online images and preventing image capture by cameras.

A. Preventing Online Image Scraping

A large portion of face images used to build today's FR systems are scraped from online social media platforms. Thus an effective way to stop $\mathcal{F}$ is to prevent web scraping. While each single user can try their best to limit their online footprint, most of the AFR methods require the help of others or an online platform (e.g., Flickr).

Anti-scraping by online platforms. *Anti-scraping* techniques have been widely studied in the security community [115], [116], [117], [118], [119]. Techniques such as rate limits, data limits, ML-based scraping detection are already used by online platforms [120]. However, a significant portion of scraping still goes undetected as scrapers develop more sophisticated tools to bypass detection [120].

Data leverage by users. $\mathcal{P}$ could try to prevent $\mathcal{F}$ from collecting their online images by withholding them. Recent works propose the concept of "data leverage" where users of online platforms work collectively to withhold data or control how their data is used by tech companies [100], [32], [121]. While not specifically aimed at facial recognition, these proposals offer alternative models for online engagement while protecting user data.

B. Avoiding Image Capture

Ordinary civilians can already use smartphones to take high-quality photos of anyone at any moment. These photos could be collected and used by facial recognition systems like PimEyes [91]. Furthermore, face photos taken by on-street surveillance cameras are increasingly used by commercial or government facial recognition systems [56], [53], [122], [1], [9], especially in major metropolitan areas and inside stores. Today's proposals for avoiding image capture come from both research community and activists (e.g. protesters and artists) concerned about surveillance. They fall into two broad categories: *hiding faces from cameras* and *disrupting camera operation*.

Face hiding. People can wear clothes, hats, masks, or move their head to prevent (usable) facial image being captured by cameras. Notably, during the June 2020 wave of protests in the US, nonprofit organizations compiled a "tech toolkit" to help privacy-conscious protesters obfuscate their faces from cameras and avoid identification [123]; in late 2020, a Chinese artist used a map of on-street surveillance cameras to successfully guide others to evade identification by positioning their head/body "away" from those cameras [124].

Camera disruption. Without physically breaking cameras, human users can prevent cameras from capturing (usable) images by simply shining laser lights at them [93]. Other commonplace methods include covering cameras with fabric or stickers.

VI. Attacking ② to Disrupt Face Pre-processing

Stage ② processes raw face images using a series of digital transformations to facilitate further operations in stages ③, ④, and ⑤. AFR proposals targeting this stage seek to disrupt the digital transformation process such that the processed face images are "unusable" by subsequent stages.

- **$\mathcal{F}$'s goal** is to obtain well-structured face images from a large number of raw images.
- **$\mathcal{P}$'s goal** is to either prevent their face being detected/extracted from raw images or to anonymize their face in these images.

A. Preventing Face Detection ②a

Face detection extracts well-centered head shots from raw images. The commonly used face detection systems [62] rely on DNNs to accurately infer the location of faces in an image. To disrupt face detection, existing AFR tools leverage the concept of "adversarial perturbations" against DNN models. Adversarial perturbations are a well-studied phenomenon in the field of adversarial machine learning. These carefully crafted, pixel-based perturbations, when added to an image, can cause DNNs to produce wrong classification results(e.g., [125], [126], [127], [128]). Typically, the perturbations are generated using an iterative optimization procedure that maximizes the likelihood of model misbehavior while minimizing perturbation visibility. The generation procedure varies depending on $\mathcal{P}$'s knowledge on $\mathcal{F}$(e.g. white-box vs black-box, see Table V).

AFR tools using adversarial perturbations can be further divided into two types, based on how the perturbation is added to images. They can be directly added to digital images if $\mathcal{P}$ has direct access to these images or fabricated as physical objects that $\mathcal{P}$ can carry (e.g., an adversarial T-shirt) or place on cameras.

Directly modifying digital images. Using AFR tools, users who post images online can directly add adversarial perturbations to these images before posting them (*e.g.*, [28], [25]). In this way, users can ensure that those properly perturbed images cannot be used by FR systems to extract any face information.

Wearing custom designed physical objects. Often users do not have access to face images to modify them. An alternative way to "inject" adversarial perturbations into images is to carry or wear a physical object so that any camera taking a photo of the user will also capture a version of the adversarial perturbation. Along these lines, prior works have successfully translated face-detection-evading adversarial perturbations into makeup [33], [123], t-shirts [26], [24], or stickers.

Placing a sticker on cameras. An orthogonal approach involves transforming the adversarial perturbation into a translucent sticker that can be placed over a camera lens. This sticker imperceptibly modifies images taken by the camera to prevent people and faces from being detected in those images [101].

B. Anonymizing Faces 2b

$\mathcal{P}$ can also *anonymize* their face images to remove identity information. Physical anonymization can be easily achieved by wearing masks, hats, makeup, etc, which overlaps with "avoiding image capture" in ① discussed in §V-B. Thus our discussion below focuses on *digital anonymization* techniques applied to *online face images*.

To anonymize face images, the leading proposals use generative adversarial networks (GANs) [129] and differential privacy [130]. Several proposals use GANs to first transform face images into latent space vectors, modify those vectors to remove identity information, and reconstruct the images from the modified vectors [102], [17], [103]. The modified faces still look human but are anonymized to prevent accurate identification. Another proposal, IdentityDP [18], uses similar techniques but goes a step further by providing provably differentially private identity protection.

A side effect of anonymization is that the anonymized faces generally do not resemble the original face but carry significant changes in shape, skin tone, hair color, or other properties.

VII. Attacking ③ to Corrupt Feature Extractor

All FR systems require an effective feature extractor to distinguish between faces of different people. AFR proposals attacking stage ③ focus on manipulating or corrupting the process of training feature extractors.

- $\mathcal{F}$**'s goal** is to train a high-quality feature extractor using available data.
- $\mathcal{P}$**'s goal** is to prevent their photos from being used to train an effective feature extractor.

A. Poisoning Training Data of Feature Extractor

Data poisoning is a well-studied technique in the field of adversarial machine learning. By manipulating the training data of a DNN model, an external party can negatively impact the model's training [131], [132], [133], [134], [135]. Poisoned models can exhibit a variety of (mis)behaviors, from incorrect classification of specific inputs to complete model failure. Existing AFR proposals focus on the latter.

Making training data unlearnable. By injecting specially crafted noise on training data, Huang *et al.* [20] render the data "unlearnable" by a DNN model. This noise misleads the model into thinking that the data have already been learned, thwarting necessary parameter updates. When a user submits their "unlearnable" face images as a training image for the FR feature extractor, the extractor will not learn anything to improve its performance. Since training an effective face feature extractor requires millions or even billions of face images [63], [64], [65], once the number of unlearnable training images becomes large enough, the trained feature extractor will not meet the accuracy level required for practical deployment.

Adding adversarial shortcuts. A related proposal from Evtimov *et al.* [104] injects *adversarial shortcuts* into the dataset. Models trained on this data overfit to the shortcut and fail to learn the meaningful semantic features of the data. Now the trained extractor model has a distorted understanding of the feature space, it cannot produce high quality feature vectors required for accurate face recognition.

VIII. Attacking ④ to Corrupt Reference Database

In stage ④, with a trained extractor in hand, $\mathcal{F}$ creates a reference database of labeled face feature vectors to facilitate identification of unidentified faces. AFR tools targeting this stage seek to fill the reference database with incorrect face/label mappings, so that $\mathcal{P}$ cannot be accurately recognized from their query images.

- $\mathcal{F}$**'s goal** is to create a database against which they can run facial recognition searches. This database should contain feature vectors of the people $\mathcal{F}$ wishes to recognize.

- **$\mathcal{P}$'s goal** is to disrupt the feature vector creation process. This prevents $\mathcal{F}$ from creating an accurate feature vector which can be matched against query images of $\mathcal{P}$'s face.

A. Poisoning Reference Feature Vectors

Existing AFR proposals in this category focus on poisoning feature vectors before they are stored into the reference database. The specific poisoning techniques depend on the underlying assumptions about how $\mathcal{F}$ compares run-time query images to the feature vectors stored in the database.

Assuming classification-based query matching. A recent AFR proposal, Fawkes [19], assumes that $\mathcal{F}$ produces run-time facial recognition results by adding a shallow classification layer on top of the feature extractor. Fawkes seeks to corrupt the final classification output by "cloaking" (or poisoning) reference images of $\mathcal{P}$, i.e. shifting their feature vectors away from the correct representation by adding imperceptible perturbations to the $\mathcal{P}$'s reference images [19]. $\mathcal{F}$'s shallow classification models trained on these shifted feature vectors will learn to associate incorrect feature spaces with $\mathcal{P}$'s identity, producing wrong matches for $\mathcal{P}$'s (uncloaked) query images at run-time. An earlier work, FishyFace [136], also proposes to disrupt face verification by poisoning the training data used to train a one-class SVM model. Since FishyFace targets per-user face verification, rather than large-scale FR systems, we exclude it in Table V and our analysis.

Assuming nearest neighbor-based matching. Two other AFR proposals, LowKey [22] and FoggySight [21], assume a K-nearest neighbors approach to query/database matching. LowKey [22] adds digital adversarial perturbations to change the feature representation of $\mathcal{P}$'s reference images (similar to Fawkes). These perturbed images create a reference feature vector for $\mathcal{P}$ that is different from those of $\mathcal{P}$'s run-time query images, thus preventing matching. FoggySight [21] takes a community-driven approach, where users modify their images to protect others. These collective modifications flood the top-K matching set for a specific user with incorrect feature vectors, drowning out the correct feature vector and preventing query image matching.

IX. Attacking Stage ⑤ to Evade Run-time Identification

The final set of AFR tools aims to prevent run-time query image identification. These methods can provide one-time protection for users who believe their images are already enrolled in a reference database. Furthermore, since labeled query images can also be added to the reference database, using these AFR tools at run-time can also help poison the reference feature vectors (see §VIII). However, current AFR proposals targeting this stage focus strictly on evasion and do not consider this joint evasion and poisoning possibility.

- **$\mathcal{F}$'s goal** is to identify the individual in the query image.
- **$\mathcal{P}$'s goal** is to alter their query image so it doesn't match their database feature vector and thus cannot be identified.

The assumption here is that $\mathcal{F}$'s reference database contains accurate feature vectors of $\mathcal{P}$.

A. Evading Identification via Adversarial Perturbations

Adversarial perturbations have been the dominant method for evading DNN classification and consequently are relevant for evading FR. Due to the extremely high number of these techniques, we restrict our discussion to proposals explicitly designed to evade FR systems at run-time. We organize these proposals by their operational context: physical and digital.

Physical evasion techniques. The first group of proposals injects adversarial perturbations into face images by having $\mathcal{P}$ wear them as physical objects. While these methods echo those described in §VI-A, they focus on thwarting image recognition or classification rather than face detection. Earlier proposals [106], [105] use adversarial makeup and eyeglasses to cause incorrect classification by FR models. More recent proposals consider two other directions, either using larger but input-independent adversarial patches to boost the effectiveness of evasion [30], or making the perturbation digitally controllable and/or much less perceivable by human eyes by projecting visible/infrared light onto user faces [112], [108], [29].

Digital evasion techniques. Here $\mathcal{P}$ digitally modifies their unlabeled (online) face images to prevent them from being accurately recognized by FR systems. Most proposals in this category apply traditional adversarial perturbation generation techniques to create minimally visible perturbations that cause $\mathcal{F}$'s feature extractor to produce misleading feature vectors. Their generation process varies depending on the assumption of feature matching process: a shallow classification on the feature vector or nearest neighbor based vector matching [110], [107], [109], [113].

More recent proposals propose methods designed to be more robust to real-world FR systems (i.e. joint optimization on multiple feature extractors, etc) [27], [23], [114]. Another recent proposal [111] uses a GAN to generate adversarial perturbations rather than applying the above mentioned optimization techniques.

X. Goals and Tradeoffs in the AFR Design Space

In our discussion of current AFR tools, we consider the design space of AFR tools through the lens of specific FR stages they disrupt. To date, all existing AFR proposals we analyzed have focused their design around disrupting a single stage in this framework. Assuming an AFR tool must disrupt some portion of the FR pipeline to be effective, we can map out and explore the design space of AFR tools using this framework.

For researchers and practitioners in the AFR community, perhaps the most critical question is: *"what are the benefits and limitations of AFR tools that target each specific stage in the framework?"* Or, an alternative form of the question might

Stage Targeted	**AFR Property**				
	Long-term Robustness	*Broad Coverage*	*No 3rd Party Assistance*	*Disruption to $\mathcal{P}$*	*Disruption to others*
①	○	**?**	**?**	○	●
②	○	**?**	●	○	●
③	**?**	**?**	**?**	○	**?**
④	○	○	●	○	**?**
⑤	**?**	●	●	○	**?**

TABLE VI
EVALUATING AFR TOOLS USING FIVE PROPERTIES, WHERE THE TOOLS ARE GROUPED BY THE FR STAGE THEY TARGET.

be: *"Given a set of prioritized properties for an AFR system, can I find the best stage(s) to disrupt in order to achieve them?"*

We attempt to answer these questions here, by first identifying a set of high level properties that AFR tools can potentially optimize for, then for each property, discussing how targeting a given stage affects an AFR tool's ability to achieve it. Ultimately, we hope to provide a high level roadmap that can guide the design of AFR tools optimizing for specific properties in mind. Note that while we consider each stage in isolation, it might be possible for an AFR tool to target multiple stages, possibly gaining a combination of benefits (and limitations).

A. Five Properties to Consider for AFR Design

When considering properties to guide the design of AFR tools, we assume that efficacy is a given priority. Our list of 5 properties target additional considerations beyond basic efficacy, and include desirable properties for efficacy (#1 and #2) and for minimizing dependencies and cost (#3, #4, #5):

1) *Long-term robustness* against evolving FR systems
2) *Broad protection coverage*, efficacy even for users with unprotected face images online
3) *No reliance on 3rd parties*, does strong protection require assistance from service providers or other users?
4) *Minimal friction for user $\mathcal{P}$*, minimizing cost for user to deploy the AFR tool on a consistent basis
5) *Minimal impact on other users*, minimizing potential risks to non-users of the AFR tool

B. Implications of AFR Designs on Key Properties

Next, we discuss the above properties in turn, and consider how easily each property can be achieved by AFR tools that target different operational stages in our framework.

For each combination of property and target stage, we "quantify" how easily the desirable property can be achieved by an AFR tool designed to disrupt that stage. ● means that the property has already been achieved by current AFR proposals targeting this stage; ○ means that the property seems "promising" and has good potential to be achieved by AFR designs targeting this stage; and **?** indicates significant progress may be required to achieve this property by targeting this stage, and the likelihood of success is unknown.

Table VI provides an overview of our conclusions. For easy notation, we will use **AFRⓚ** to refer to the group of AFR proposals that target FR stage ⓚ.

Property 1: Long-term robustness

An effective AFR tool should provide strong and lasting protection against unwanted facial recognition. That is, it should protect a user $\mathcal{P}$ from unwanted FR from initial use, and extending into the future, even as FR systems continue to advance.

●: None While this principle is the main goal of AFR, none of existing AFR tools (targeting any stage) is able to achieve this property. No current system provides strong protection against ever-evolving FR systems.

○: AFR①, AFR②, AFR④ Conceptually, $\mathcal{P}$ can achieve long-term robustness by consistently undermining the face data pipeline of $\mathcal{F}$. AFR① and AFR② can both prevent any face image of $\mathcal{P}$ to be included into $\mathcal{F}$'s pipeline. AFR④ can corrupt $\mathcal{F}$'s understanding of any face images in the reference database. While promising, existing AFR tools fail to *consistently* prevent the inclusion of or corrupt *all* $\mathcal{P}$'s images from both online and physical sources.

?: AFR③, AFR⑤ It remains unclear if these two groups of AFR tools can provide long-term robustness. AFR③ could be overcome over time as $\mathcal{F}$ switches to newer and different feature extractors. AFR⑤ offers only one-time protection, and does not address the scenario where query images get added to the reference database.

Property 2: Broad protection coverage

Many of us already have an online presence, e.g., face photos posted years ago without AFR protection. An effective AFR proposal would ideally provide protection under the challenging but realistic scenario where $\mathcal{P}$ already has unprotected face images online.

●: AFR⑤ AFR tools that rely on run-time evasion are not impacted by the existence of unprotected images online.

○: AFR④ The presence of unprotected images complicates the protection of AFR④ since $\mathcal{F}$ has some groundtruth information about the correct features of $\mathcal{P}$'s faces. However, the addition of protected images can slowly move the features of $\mathcal{P}$ away from the correct feature, and thus achieve protection. Moreover, several AFR tools [19], [21] proposed a "group cloaking" idea where multiple users coordinate together to achieve better protection for those having an existing online presence.

?: AFR①, AFR②, AFR③ These three groups of AFR tools focus on disrupting the (training) data pipeline of FR. As a result, they cannot protect $\mathcal{P}$ against $\mathcal{F}$ who has obtained unprotected images of $\mathcal{P}$.

Property 3: No reliance on 3rd party to operate

Ideally, an AFR tool can be operated by a user $\mathcal{P}$ alone, and achieve strong protection without assistance or participation third-party, either a central content provider like Facebook or a friendly user willing to cooperate to help $\mathcal{P}$. This is an abstract measure of the entity-level complexity required to operate the tool. Achieving this property has the added benefit of limiting exposure of potentially sensitive user photos or personal data to any 3rd party, *i.e.* the AFR is also privacy-preserving.

●: AFR②, AFR④, AFR⑤ AFR tools in these three groups all rely on adding certain perturbations on face images, which can be done by $\mathcal{P}$ without assistance from other parties.

?: AFR①, AFR③ For those AFR① seeking to prevent online data scraping, they rely on the assistance of image sharing platforms. Similarly, disrupting the training of a feature extractor requires a coordinated effort across many users, since $\mathcal{P}$ only contributes a very limited subset of the training data.

Property 4: Minimal disruption to $\mathcal{P}$

This usability-related property measures what $\mathcal{P}$ needs to sacrifice in order to consistently apply the AFR tool. This property is motivated by the well-known findings that users prefer and are more likely to use protection solutions that introduce minimal friction to their daily life [137], [138].

○: AFR①, AFR②, AFR③, AFR④, AFR⑤ So far, existing AFR tools all introduce some levels of "disruption" to $\mathcal{P}$, whether it is adding visual noise, perturbations or transformations to $\mathcal{P}$'s online photos that rampages their original purpose, requiring $\mathcal{P}$ to always wear odd makeup/clothes/accessories, or purchasing more powerful computing hardware/services to implement the AFR tool against continuely evolving $\mathcal{F}$. More research efforts are needed to limit the amount/type of disruption to users.

Property 5: Minimal impact on other users

This final property examines how the outcome of $\mathcal{P}$'s AFP protection would affect other users. Intuitively, $\mathcal{P}$ can protect themselves by forcing $\mathcal{F}$ to fail (give a null or uninformative result), or by intentionally tricking $\mathcal{F}$ to recognize them as another person $\mathcal{P}$'. Depending on the context, the latter may negatively affect $\mathcal{P}$', producing potential social risks (see §XI-B for detailed discussions on social challenges facing AFR).

●: AFR①, AFR② These two groups of AFR tools focus on disrupting the data pipeline of $\mathcal{F}$, and thus, have no impact on other users.

?: AFR③, AFR④, AFR⑤ These three groups of AFR tools seek to intentionally misclassify $\mathcal{P}$'s face to another user, and as a result, could potentially impact other users included in $\mathcal{F}$'s reference database.

XI. Challenges for AFR Tools

In this section, we describe what we see as the major technical and broader social/ethical challenges facing future AFR development. Each challenge spans multiple properties and stages laid out in this paper. For each challenge, we provide context for why the challenge exists and, where possible, suggest ways to address it. Like §X, the challenges described here represent our best efforts to understand and systematize the AFR space. They are not exhaustive, and are meant as signposts rather than a comprehensive map for future research.

A. Technical Challenges

TC 1: Lack of provable protection

Our analysis shows that the majority of AFR proposals, especially those targeting stages ② – ⑤, employ adversarial perturbations, which do not yet provide provable protection guarantees. In practice, the success rate of adversarial perturbations may drop significantly when $\mathcal{P}$'s knowledge of $\mathcal{F}$ is imperfect [139]. Many adversarial perturbation-based protections can also be circumvented by more advanced FR systems. For example, $\mathcal{F}$ could adversarially train the feature extractor [140], [141] to be more robust against adversarial examples, thus defeating AFR tools against stages ③ or ④. $\mathcal{F}$ could also remove adversarial perturbations from face images before processing them or adding them to the reference database [142], circumventing AFR tools that target stages ② or ⑤.

Potential Directions. Improving adversarial perturbation generation methods may help increase short-term efficacy of those AFR tools. However, the lack of provable, ongoing protection is a much tougher barrier to overcome. In order to provide reliable, ongoing protection, developers of AFR tools can consider two possible paths: (i) integrate provable guarantees into the perturbation generation process, or (ii) consider an alternative that provides guaranteed protection. For (ii), there are two potential directions. The first is focus on attacking stage ①, where defeating FR does not require evading or poisoning a feature extractor. The second is to switch from "misleading" the feature extractor with "minor" image modifications to completely disabling the feature extraction and/or matching process.

TC 2: Existence of online footprints

Some AFR proposals (especially those targeting stage ④) implicitly or explicitly assume that users can start "from scratch" to protect their online persona. In practice, most Internet users today already have face images online, posted by themselves or others, and at least some of those images are already captured by FR databases. Over 1.8 billion photos are uploaded to online platforms daily [143], making it likely that one or more unmodified photos of a user $\mathcal{P}$ will likely end up online, with or without $\mathcal{P}$'s knowledge. Given the widespread use of web scraping to collect FR reference images [91], [5], it is likely that at least one of these photos is already in a FR system reference database.

Potential Directions. This stark reality has two implications for future AFR research. *First*, AFR tools should be evaluated under the practical scenarios where the FR system has access to both protected and unprotected online photos of $\mathcal{P}$. While several AFR tools have provided such measurements (*e.g.* [19], [21]), many others have not. *Second*, we believe that AFR tools managed by online platforms will offer better protection of online footprints against FR systems than those executed by individual users. These platforms can protect photos of an individual posted by them or others, and are overall better positioned to deploy more powerful protection mechanisms.

For example, online platforms could employ the group cloaking techniques proposed in Fawkes [19] or FoggySight [21] to corrupt reference databases composed of images from their sites. After images are scraped, online platforms could use provenance-tracking to re-identify stolen images, *e.g.* in the training dataset of a feature extractor, and enable exposure/prosecution of photo thieves [144], [145], [146]. All these methods ought to be accompanied by enhanced anti-scraping techniques to prevent large-scale scraping of face images, i.e. stricter rate limiting, access permissions, and scraping detection heuristics, to make it safer for individuals to have online footprints.

TC 3: Face images don't change

A related but distinct challenge faced by AFR systems is the permanence of face data. For better or for worse, most people have the same face their whole adult life [5]. Our faces may age, but they remain recognizable as uniquely "us" to most humans and FR systems [147]. The slow rate at which faces change is a major challenge for AFR tools. To be long-term effective, these tools must conceal the same piece of static data (a face) from numerous adversaries over many years.

Once $\mathcal{F}$ obtains $\mathcal{P}$'s protected face photo, they can try as many times as they want to break the protection [141]. If $\mathcal{F}$ ever succeeds, either in 1 month or 1 year, they "win" and $\mathcal{P}$ loses, because modern FR systems only need one clean picture in the reference database to identify a person [63]. For example, Clearview.ai identified a person based on a single reference image in which the person's reflection appeared faintly in a mirror [5]. Clearly, the issue of face data permanence poses a significant challenge for AFR tool development.

TC 4: Lack of transparency of FR systems

One final technical challenge faced by AFR tool developers is the lack of transparency on how proprietary FR systems work in practice. This hampers AFR tool development and testing. Without access to proprietary FR systems, AFR researchers must do their best to glean a generic understanding of how FR systems work from public documents and academic papers, e.g. [65], [68]. While this may be sufficient to develop AFR tools that work well in the lab, it would likely be impossible for researchers to perform comprehensive efficacy tests against proprietary systems.

Furthermore, AFR tool developers have no knowledge of how or if FR systems are actively working to overcome AFR systems. The 2020 global FR market was valued at 3.86 billion US dollars [148], so FR stakeholders have ample resources and personnel to quickly deploy changes as new AFR systems emerge. Even passive improvements to FR systems, such as the arrival of new training methods or architectures, can overcome AFR protection and compromise user privacy [141]. Altogether, this lack of transparency means that that AFR tools face an upward battle in the fight against unwanted FR.

B. Broader Social and Ethical Considerations

In addition to these technical challenges, AFR tools face broader social and ethical considerations. These stem from a variety of factors, including a lack of regulation, benefits of FR for the public good, and demographic disparities in FR systems.

SC 1: Unregulated, ubiquitous FR

Today, FR systems are generally unregulated and easy to deploy. Practically anyone with a powerful laptop and access to an image dataset could create a FR system. This democratization of FR has allowed 3rd party FR systems like Clearview.ai, which rely on unauthorized data use [61], to flourish. As a result, it is extremely difficult (if not impossible) for individuals to know when/where FR systems are deployed and what they are capable of.

This lassiez-faire climate creates significant ambiguity as to when AFR tools can/should be deployed. For example, around the world, photos taken for official government purposes (*e.g.* drivers' license and passport photos) are used as reference images in government FR systems aiding law enforcement officers, border control agents, among others [1], [48], [50], [3]. This goverment-sponsored FR may be *unwanted* but is not (necessarily) *unauthorized* under the status quo, and the legality of using AFR tools to thwart downstream FR when official driver's license photos are taken is ambiguous. To augment the confusion, systems like Clearview are used by law enforcement [5], further blurring the concept of *unauthorized* vs *unwanted* FR and the appropriate use of AFR tools. As FR and AFR use increases, a clash over this issue seems almost inevitable.

[5] Major plastic or reconstructive surgery excepted.

SC 2: FR used for social good

Both privacy-sensitive citizens and criminals can use AFR tools. Law enforcement's use of facial recognition can benefit society in multiple ways, such as tracking and locating wanted criminals or lost children [149], [150]. Consequently, AFR tools applied by bad actors could ultimately harm the public good. The debate between privacy and national security plays out in numerous other tech domains, such as end-to-end encryption [151]. Legitimate claims can be made by both sides. AFR researchers must be mindful of this tension and the potential consequences of their work.

SC 3: Harm caused by AFR misidentification

One ethical tension not yet explored in current literature is the social effect of misidentifications caused by AFR tools. For example, if U uses an AFR tool and is misidentified by a recognition system as P, what outcome might this have for P? If U is engaging in illegal activity but P is arrested instead, the AFR tool could cause serious harm, both to P and to U's victim(s).

The well-known bias of FR systems heightens this tension. Police departments routinely make rushed identification decisions based on partial results from facial recognition systems [81]. Furthermore, facial recognition systems misidentify people of color at higher rates [34], [152]. Recent work has found that AFR tools exhibit these same biases [153], [154]. The social impact of AFR misclassification requires urgent study.

XII. Concluding Thoughts

As facial recognition (FR) continues to grow in scale and ubiquity, we expect anti-facial recognitions to rise in popularity. There is an urgent need to think longitudinally about AFR tools, analyzing both their limits and their potential. Our paper aims to fill this gap by providing both a framework for discussing AFR proposals and an assessment of the current state of AFR research.

We find that current AFR tools possess some, but not all, of the traits needed to successfully defeat unwanted FR in the real world. Many existing proposals leverage adversarial perturbations to evade FR models, either in the preprocessing ② or classification ⑤ stages. Such perturbations, while often effective in the short-term, lack long-term guarantees, and cannot fundamentally change FR system behavior in the future. Future AFR proposals may benefit from more exploration of designs that target stages ① and ④, which could provide wider-reaching protection.

References

[1] C. Garvie, A. Bedoya, and J. Frankle, "The perpetual lineup," *Georgetown Law Center on Privacy and Technology*, 2016.

[2] J. Chin and C. Burge, "Twelve days in xinjiang: How china's surveillance state overwhelms daily life," *The Wall Street Journal*, 2017.

[3] P. Mozur, "One month, 500,000 face scans: How china is using ai to profile a minority," *The New York Times*, vol. 14, 2019.

[4] P. Reevell, "How russia is using facial recognition to police its coronavirus lockdown," April 30, 2021. [Online]. Available: https://abcnews.go.com/International/russia-facial-recognition-police-coronavirus-lockdown/story?id=70299736

[5] K. Hill, "The secretative company that may end privacy as we know it," *The New York Times*, 2020.

[6] ——, "Clearview ai's facial recognition app called illegal in canada," *The New York Times*, 2021.

[7] "Collection of biometric data from aliens upon entry to and departure from the united states." [Online]. Available: https://www.regulations.gov/docket/USCBP-2020-0062/document

[8] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing face recognition vendor test (frvt)," *NIST*, 2018. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8238.pdf

[9] C. Garvie and L. Moy, "America under watch," *Georgetown Law Center on Privacy and Technology*, 2019.

[10] Y. Pan, Q. Feng, and C. Zhang, "Face recognition at the sales office," *AI Outpost*, 2020. [Online]. Available: https://mp.weixin.qq.com/s/fWbQ3SD9vB-QdB51T097hw

[11] "Fears for children's privacy as delhi schools install facial recognition," *Reuters*, 2021. [Online]. Available: https://www.reuters.com/article/us-india-tech-facialrecognition-trfn-idUSKBN2AU0P5

[12] D. Jeans, "Amazon extends moratorium on police use of facial recognition technology," *Forbes*, 2020.

[13] A. Krishna, "Ibm ceo's letter to congress on racial justice reform," 2020.

[14] "Ban dangerous facial recognition technology that amplifies racist policing," *Amnesty International*, 2021.

[15] "Stop facial recognition," *Big Brother Watch*, 2021. [Online]. Available: https://bigbrotherwatch.org.uk/campaigns/stop-facial-recognition/

[16] S. Rodriguez, "Facebook plans to shut down its facial recognition program," 2021. [Online]. Available: https://www.cnbc.com/2021/11/02/facebook-will-shut-down-program-that-automatically-recognizes-people-in-photos-and-videos-delete-data.html

[17] T. Li and M. S. Choi, "Deepblur: A simple and effective method for natural image obfuscation," *arXiv preprint arXiv:2104.02655*, vol. 1, 2021.

[18] Y. Wen, L. Song, B. Liu, M. Ding, and R. Xie, "Identitydp: Differential private identification protection for face images," *arXiv preprint arXiv:2103.01745*, 2021.

[19] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting privacy against unauthorized deep learning models," in *Proc. of USENIX Security*, 2020.

[20] H. Huang, X. Ma, S. M. Erfani, J. Bailey, and Y. Wang, "Unlearnable examples: Making personal data unexploitable," in *Proc. of ICLR*, 2021.

[21] I. Evtimov, P. Sturmfels, and T. Kohno, "Foggysight: a scheme for facial lookup privacy," *Proc. of PETS*, 2021.

[22] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," in *Proc. of ICLR*, 2021.

[23] T. Cilloni, W. Wang, C. Walter, and C. Fleming, "Preventing personal data theft in images with adversarial ml," *arXiv preprint arXiv:2010.10242*, 2020.

[24] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, "Making an invisibility cloak: Real world adversarial attacks on object detectors," in *Proc. of ECCV*, 2020.

[25] M. Treu, T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Fashion-guided adversarial attack on person segmentation," in *Proc. of CVPR*, 2021.

[26] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in *Proc. of ECCV*, 2020.

[27] V. Chandrasekaran, C. Gao, B. Tang, K. Fawaz, S. Jha, and S. Banerjee, "Face-off: Adversarial face obfuscation," *Proc. of PETS*, 2021.

[28] M. Xue, S. Sun, Z. Wu, C. He, J. Wang, and W. Liu, "Socialguard: An adversarial example based privacy-preserving technique for social images," *arXiv preprint arXiv:2011.13560*, 2020.

[29] D.-L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial light projection attacks on face recognition systems: A feasibility study," in *Proc. of CVPR*, 2020.

[30] S. Komkov and A. Petiushko, "Advhat: Real-world adversarial attack on arcface face id system," in *Proc. of ICPR*. IEEE, 2021.

[31] K. Browne, B. Swift, and T. Nurmikko-Fuller, “Camera adversaria,” in *Proc. of CHI*, 2020, pp. 1–9.

[32] N. Vincent and B. Hecht, “Can “conscious data contribution” help users to exert “data leverage” against technology companies?” *Proc. of CHI*, 2021.

[33] A. Harvey, “Cv dazzle: Camouflage from face detection,” *Master’s thesis*, 2010.

[34] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Proc. of FaaCT*, 2018.

[35] M. Taskiran, N. Kahraman, and C. E. Erdem, “Face recognition: Past, present and future (a review),” *Digital Signal Processing*, p. 102809, 2020.

[36] L. Mackenzie, “Surveillance state: how gulf governments keep watch on us,” *Wired Magazine*, 2021. [Online]. Available: https://wired.me/technology/privacy/surveillance-gulf-states/

[37] L. Kayali, “How facial recognition is taking over a french city,” 2019. [Online]. Available: https://www.politico.eu/article/how-facial-recognition-is-taking-over-a-french-riviera-city/

[38] C. Burt, “Kenyan police launch facial recognition on urban cctv network,” 2018. [Online]. Available: https://www.biometricupdate.com/201809/kenyan-police-launch-facial-recognition-on-urban-cctv-network

[39] K. Pivcevic, “Police facial recognition use in belarus, greece, myanmar raises rights, data privacy concerns,” *Biometric Update*, 2021. [Online]. Available: https://www.biometricupdate.com/202103/police-facial-recognition-use-in-belarus-greece-myanmar-raises-rights-data-privacy-concerns

[40] P. Fussey and D. Murray, “Independent report on the london metropolitan police service’s trial of live facial recognition technology,” 2019.

[41] S. Jie, “China exports facial id technology to zimbabwe,” *Global Times*, vol. 12, 2018.

[42] R. Mellen, “Buenos aires is using facial recognition system that tracks child suspects, rights group says,” 2020. [Online]. Available: https://www.washingtonpost.com/world/2020/10/09/argentina-facial-recognition-juvenile-suspects/

[43] H. Devlin, “We are hurtling towards a surveillance state: the rise of facial recognition technology,” 2019. [Online]. Available: https://www.theguardian.com/technology/2019/oct/05/facial-recognition-technology-hurtling-towards-surveillance-state

[44] P. Mozur, “Inside china’s dystopian dreams: A.i., shame and lots of cameras,” *The New York Times*, 2018.

[45] C. Tan, “Malaysian police adopt chinese ai surveillance technology,” *Nikkei Asia*, 2018. [Online]. Available: https://asia.nikkei.com/Business/Companies/Chinas-startup-supplies-AI-backed-wearable-cameras-to-Malaysian-police

[46] A. Mascillno, “Facial recognition in schools: systems deployed in europe and the us amid privacy concerns,” 2021. [Online]. Available: https://www.biometricupdate.com/202110/facial-recognition-in-schools-systems-deployed-in-europe-and-the-us-amid-privacy-concerns

[47] M. Andrejevic and N. Selwyn, “Facial recognition technology in schools: Critical questions and concerns,” *Learning, Media and Technology*, vol. 45, no. 2, pp. 115–128, 2020.

[48] A. Ziv, “This israeli face-recognition startup is secretly tracking palestinians,” 2019. [Online]. Available: https://www.haaretz.com/israel-news/business/.premium-this-israeli-face-recognition-startup-is-secretly-tracking-palestinians-1.7500359

[49] M. Z. Khan, “System soon to identify risky goods, individuals at borders,” *Dawn*, 2020. [Online]. Available: https://www.dawn.com/news/1584264

[50] M. Mason, “Biometric breakthrough: How cbp is meeting its mandate and keeping america safe,” *U.S. Customs and Border Protection Website*. [Online]. Available: https://www.cbp.gov/frontline/cbp-biometric-testing

[51] “Facial recognition tech fights coronavirus in chinese city,” *France24*, 2020. [Online]. Available: https://www.france24.com/en/live-news/20210713-facial-recognition-tech-fights-coronavirus-in-chinese-city

[52] A. Roussi, “Resisting the rise of facial recognition,” 2021. [Online]. Available: https://www.nature.com/articles/d41586-020-03188-2

[53] T. Clayburn, “Apple sued in nightmare case involving teen wrongly accused of shoplifting, driver’s permit used by impostor, and unreliable facial-rec tech,” *The Register*, 2021.

[54] “The retail stores you probably shop at that use facial-recognition technology.” [Online]. Available: https://www.businessinsider.com/retail-stores-that-use-facial-recognition-technology-macys-2021-7

[55] U. Saiidi, “We went inside alibaba’s global headquarters. here’s what we saw,” *CNBC*, 2019. [Online]. Available: https://www.cnbc.com/2019/09/11/we-went-inside-alibabas-global-headquarters-heres-what-we-saw.html

[56] M. Rogoway, “Major tech company using facial recognition to id workers,” *The Oregonian*, 2020. [Online]. Available: https://www.govtech.com/public-safety/major-tech-company-using-facial-recognition-to-id-workers.html

[57] “In-car biometric technology for human interaction,” *Hyunai Motor Group*, 2020. [Online]. Available: https://tech.hyundaimotorgroup.com/article/in-car-biometric-technology-for-human-interaction/

[58] P. Lyon, “Subaru forester is first mainstream model to offer facial recognition technology,” *Forbes*, 2018. [Online]. Available: https://www.forbes.com/sites/peterlyon/2018/04/30/subaru-forester-is-first-affordable-car-to-get-facial-recognition/?sh=2b939c4c569e

[59] A. Smith, “Jetblue will test facial recognition for boarding,” *CNN Business*, 2017. [Online]. Available: https://money.cnn.com/2017/05/31/technology/jetblue-facial-recognition/index.html

[60] “Could facial recognition be the future of airport security? delta air lines is testing it out,” *CBS News*, 2021. [Online]. Available: https://www.cbsnews.com/news/facial-recognition-delta-tsa/

[61] R. Heilweil, “The world’s scariest facial recognition company, explained,” *Vox.com*.

[62] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[63] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proc. of CVPR*, 2019, pp. 4690–4699.

[64] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “Cosface: Large margin cosine loss for deep face recognition,” in *Proc. of CVPR*, 2018, pp. 5265–5274.

[65] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proc. of CVPR*, 2015, pp. 815–823.

[66] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, “Magface: A universal representation for face recognition and quality assessment,” *arXiv preprint arXiv:2103.06627*, 2021.

[67] “Amazon rekognition.” [Online]. Available: https://aws.amazon.com/rekognition/customers/

[68] “Facial recognition technology: Privacy and accuracy issues related to commericial uses,” 2020.

[69] GAO-21-518, “Facial recognition technology: Federal law enforcement agencies should better assess privacy and other risks,” 2021.

[70] “Idemia.” [Online]. Available: https://na.idemia.com/

[71] “Clearview.ai.” [Online]. Available: https://clearview.ai

[72] “Azure face recognition.” [Online]. Available: https://azure.microsoft.com/en-us/services/cognitive-services/face/

[73] “Amazon rekognition.” [Online]. Available: https://aws.amazon.com/rekognition

[74] “Megvii.” [Online]. Available: https://en.megvii.com

[75] “Sensetime.” [Online]. Available: https://sensetime.com

[76] “Yitu.” [Online]. Available: https://yitutech.com/en

[77] “Cloudwalk.” [Online]. Available: https://cloudwalk.com/en

[78] C. Petters, “Build your own face recognition service using amazon rekognition,” *AWS Machine Learning Blog*.

[79] H. Jacobs and P. Ralph, “Inside the creepy and impressive startup funded by the chinese government that is developing ai that can recognize anyone, anywere,” *Business Insider*. [Online]. Available: https://www.businessinsider.com/china-facial-recognition-tech-company-megvii-faceplusplus-2018-5

[80] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning (2016),” *arXiv preprint arXiv:1602.07261*, 2016.

[81] C. Garvie, “Garbage in, garbage out: Face recognition on flawed data,” *Georgetown Law Center on Privacy and Technology*.

[82] “Huawei/megvii uyghur alarms,” 2020. [Online]. Available: https://ipvm.com/reports/huawei-megvii-uygur

[83] T. May and A. C. Chien, "Game over: Chinese company deploys facial recognition to limit youths' play," 2021. [Online]. Available: https://www.nytimes.com/2021/07/08/business/video-game-facial-recognition-tencent.html

[84] "Mapped: The state of facial recognition around the world," 2020. [Online]. Available: https://www.visualcapitalist.com/facial-recognition-world-map/

[85] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," 2016.

[86] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. of FG*, 2018, pp. 67–74.

[87] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *Proc. of ICIP*, 2014, pp. 343–347.

[88] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014.

[89] "China state tv exposes wide illegal use of facial recognition cameras in commercial properties," *China Money Network*, 2021. [Online]. Available: https://www.chinamoneynetwork.com/2021/03/16/china-state-tv-exposes-wide-illegal-use-of-facial-recognition-cameras-in-commercial-properties

[90] K. Hill, "The secretive company that might end privacy as we know it," *The New York Times*, vol. 18, p. 2020, 2020.

[91] "Pimeyes." [Online]. Available: https://pimeyes.com/en

[92] D. Byler, "Because there were cameras, i didn't ask any questions," 2020. [Online]. Available: https://www.chinafile.com/extensive-surveillance-china

[93] "Hong kong protesters are using lasers to distract and confuse. police are shining lights right back." *The Washington Post*, 2020. [Online]. Available: https://www.washingtonpost.com/world/2019/08/01/hong-kong-protesters-are-using-lasers-distract-confuse-police-are-pointing-them-right-back/

[94] B. Friedman and A. Ferguson, "Here's a way forward on facial recognition," *The New York Times*, 2019.

[95] M. Devich-Cyril, "Defund facial recognition," *The Atlantic*, 2020.

[96] A. Smith, "More than half of u.s. adults trust law enforcement to use facial recognition responsibly," *Pew Research Center*, 2019.

[97] A. L. Institute, "Beyond face value: Public attitudes to facial recognition technology," 2019.

[98] L. Steinacker, M. Meckel, G. Kostka, and D. Borth, "Facial recognition: A cross-national survey on public acceptance, privacy, and discrimination," *Proc. of ICML LML Workshop*, 2020.

[99] X. Lai and P.-L. P. Rau, "Has facial recognition technology been misused? a user perception model of facial recognition scenarios," *Computers in Human Behavior*, 2021.

[100] N. Vincent, H. Li, N. Tilly, S. Chancellor, and B. Hecht, "Data leverage: A framework for empowering the public in its relations hip with technology companies," in *Proc. of FAccT*, 2021.

[101] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," *arXiv preprint arXiv:2012.12528*, 2020.

[102] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," in *Proc. of ISVC*, 2019.

[103] S. Yang, W. Wang, Y. Cheng, and J. Dong, "A systematical solution for face de-identification," in *Chinese Conference on Biometric Recognition*. Springer, 2021, pp. 20–30.

[104] I. Evtimov, I. Covert, A. Kusupati, and T. Kohno, "Disrupting model training with adversarial shortcuts," *arXiv preprint arXiv:2106.06654*, 2021.

[105] R. Feng and B. Prabhakaran, "Facilitating fashion camouflage art," in *Proc. of ACM Multimedia*, 2013.

[106] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. of CCS*, 2016.

[107] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *Proc. of WACV*. IEEE, 2019.

[108] Z. Zhou, D. Tang, X. Wang, W. Han, X. Liu, and K. Zhang, "Invisible mask: Practical attacks on face recognition with infrared," *arXiv preprint arXiv:1803.04683*, 2018.

[109] Y. Dong, H. Su, B. Wu, Z. Li, W. Liu, T. Zhang, and J. Zhu, "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. of CVPR*, 2019, pp. 7714–7722.

[110] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, "Generating adversarial examples by makeup attacks on face recognition," in *Proc. of ICIP*. IEEE, 2019.

[111] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *Proc. of IJCB*. IEEE, 2019.

[112] M. Shen, Z. Liao, L. Zhu, K. Xu, and X. Du, "Vla: A practical visible light-based attack on face recognition systems in physical world," *Proc. of IMWUT*, 2019.

[113] I. Singh, S. Momiyama, K. Kakizaki, and T. Araki, "On brightness agnostic adversarial examples against face recognition systems," in *BIOSIG*. IEEE, 2021.

[114] X. Yang, Y. Dong, T. Pang, H. Su, J. Zhu, Y. Chen, and H. Xue, "Towards face encryption by generating adversarial identity masks," in *ICCV*, 2021.

[115] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection," in *Proc. of USENIX Security*, 2013, pp. 241–256.

[116] Z. Gold and M. Latonero, "Robots welcome: Ethical and legal considerations for web crawling and scraping," *Wash. JL Tech. & Arts*, vol. 13, p. 275, 2017.

[117] K. Parikh, D. Singh, D. Yadav, and M. Rathod, "Detection of web scraping using machine learning," *Open access international journal of Science and Engineering*, pp. 114–118, 2018.

[118] D. Jawad, "Detection of web api content scraping: An empirical study of machine learning algorithms," 2017.

[119] A. Haque and S. Singh, "Anti-scraping application development," in *Proc. of ICACCI*. IEEE, 2015.

[120] M. Clark, "Scraping by the numbers," 2021. [Online]. Available: https://about.fb.com/news/2021/05/scraping-by-the-numbers/

[121] N. Vincent, B. Hecht, and S. Sen, """data strikes": Evaluating the effectiveness of a new form of collective action against technology companies," in *Proc. of WWW*, 2019.

[122] J. J. Roberts, "Walmart's use of sci-fi tech to spot shoplifters raises privacy questions," *Fortune*, 2015. [Online]. Available: https://fortune.com/2015/11/09/wal-mart-facial-recognition/

[123] "How to protect your phone and identity at protests," *Amnesty International*, 2021. [Online]. Available: https://banthescan.amnesty.org/wp-content/uploads/2021/01/Amnesty-Tech-Toolkit-FINAL-1.pdf

[124] V. Ni and Y. Wang, "How to 'disappear' on happiness avenue in beijing," *BBC*, 2020. [Online]. Available: https://www.bbc.com/news/technology-55053978

[125] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[126] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. of IEEE S&P*, 2017.

[127] P.-Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C.-J. Hsieh, "Ead: elastic-net attacks to deep neural networks via adversarial examples," in *Proc. of AAAI*, 2018.

[128] J. Bao, "Sparse adversarial attack to object detection," *arXiv preprint arXiv:2012.13692*, 2020.

[129] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Proc. of NeurIPS*, 2014.

[130] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, 2014.

[131] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," in *Proc. of Machine Learning and Computer Security Workshop*, 2017.

[132] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[133] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. of NDSS*, 2018.

[134] C. Zhu, W. R. Huang, A. Shafahi, H. Li, G. Taylor, C. Studer, and T. Goldstein, "Transferable clean-label poisoning attacks on deep neural nets," in *Proc. of ICML*, 2019.

[135] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Proc. of NeurIPS*, 2018.

[136] G. Garofalo, V. Rimmer, D. Preuveneers, W. Joosen *et al.*, "Fishy faces: Crafting adversarial images to poison face authentication," in *Proc. of WOOT*, 2018.

[137] K. P. Coopamootoo, “Usage patterns of privacy-enhancing technologies,” in *Proc. of CCS*, 2020.

[138] R. N. Wright, L. J. Camp, I. Goldberg, R. L. Rivest, and G. Wood, “Privacy tradeoffs: Myth or reality?” in *Proc. of FC*. Springer, 2002.

[139] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli, “Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks,” in *Proc. of USENIX*, 2019.

[140] L. Chen, H. Wang, B. Z. H. Zhao, M. Xue, and H. Qian, “Oriole: Thwarting privacy against trustworthy deep learning models,” *arXiv preprint arXiv:2102.11502*, 2021.

[141] E. Radiya-Dixit and F. Tramer, “Data poisoning won’t save you from facial recognition,” *arXiv*, 2021.

[142] D. Deb, X. Liu, and A. K. Jain, “Faceguard: A self-supervised defense against adversarial face images,” *arXiv preprint arXiv:2011.14218*, 2020.

[143] P. Suciu, “A photo used to be worth a thousand words, but thanks to social media photos have lost their value,” *Forbes online*, 2019.

[144] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, “Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models,” *arXiv preprint arXiv:1806.01246*, 2018.

[145] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, “Membership inference attacks against machine learning models.” IEEE, 2017, pp. 3–18.

[146] A. Sablayrolles, M. Douze, C. Schmid, and H. Jegou, “Radioactive data: tracing through training,” in *Proc. of ICML*, 2020.

[147] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, “Face verification across age progression using discriminative methods,” *IEEE Transactions on Information Forensics and security*, 2009.

[148] G. Research, “Facial recognition market size, share and trends report, 2021 - 2028,” 2021. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/facial-recognition-market

[149] “https://findbiometrics.com/chinese-police-use-facial-recognition-find-child-abducted-30-years-ago-052107/,” *FindBiometrics*, 2020. [Online]. Available: https://findbiometrics.com/chinese-police-use-facial-recognition-find-child-abducted-30-years-ago-052107/

[150] “Pinellas county sheriff’s office facial recognition program,” *Pinellas County Sheriff’s Office*, 2019. [Online]. Available: https://www.documentcloud.org/documents/6586379-FACESlist-Redacted.html

[151] “International statement: End-to-end encryption and public safety,” *United States Department of Justice*, 2020. [Online]. Available: https://www.justice.gov/opa/pr/international-statement-end-end-encryption-and-public-safety

[152] S. Dooley, T. Goldstein, and J. P. Dickerson, “Robustness disparities in commercial face detection,” *arXiv preprint arXiv:2108.12508*, 2021.

[153] H. Rosenberg, B. Tang, K. Fawaz, and S. Jha, “Fairness properties of face recognition and obfuscation systems,” *arXiv preperint arXiv:2108.02707*, 2021.

[154] S. Qin, “Bias and fairness of evasion attacks in image perturbation,” *All Master’s Theses*, 2021. [Online]. Available: https://digitalcommons.cwu.edu/etd/1517