SoK: Anti-Facial Recognition Technology

Emily Wenger
University of Chicago
ewenger@uchicago.edu

Shawn Shan
University of Chicago
shawnshan@cs.uchicago.edu

Haitao Zheng
University of Chicago
htzheng@cs.uchicago.edu

Ben Y. Zhao
University of Chicago
ravenben@cs.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.**

1. Introduction

In recent years, facial recognition systems have accelerated their growth in scale and reach, and are becoming an increasingly ubiquitous part of our daily lives. As a result, the majority of citizens in the world's most populous countries are already enrolled in one or more facial recognition systems, whether they know it or not. For example, in the United States, nearly 200 million residents are already enrolled in the FBI's facial recognition database, which was built by leveraging FBI's access to driver license photos in many 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 real-world 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 confirmed 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]. Other reports have detailed how photos taken in unexpected places – airports, city streets, government buildings, schools, corporate offices – end up in facial recognition systems without subjects' knowledge or consent [1, 7, 8, 9, 10, 11].

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

AFR tools have also attracted significant attention from the research community. 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. Existing surveys [34, 35] on facial privacy issues do not consider user-centric AFR tools. They instead discuss privacy-preserving techniques that surveillance system operators could employ, a related but separate line of work to that addressed here (see §12).

In this paper, we address this need by providing a common framework for analyzing a wide range of AFR systems. More specifically, we make the following contributions:

- **Taxonomy of targets in facial recognition systems:** AFR systems target a wide range of components in the

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

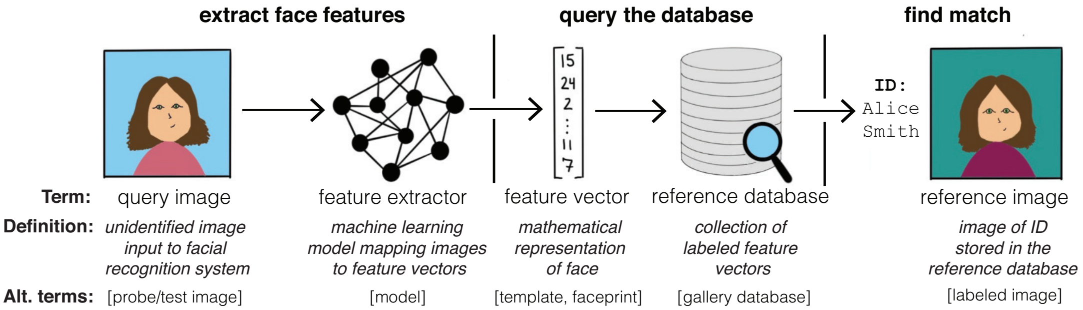

Figure 1. The workflow of how facial recognition systems recognize a human face in an input image, along with the corresponding terminology. (a): A query image, after being submitted to the system, is passed to the feature extractor to produce a feature vector; (b): this feature vector is used to query 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).

facial recognition process. Using a generalized version of the FR data pipeline, we provide a framework for reasoning broadly about existing and future AFR work.

- **Categorization and analysis of AFR tools:** We take the current body of work on AFR tools, and 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 significant challenges facing current AFR systems, as well as directions for potential solutions.

The rest of the paper proceeds as follows. We begin by providing operational details of real-world facial recognition systems (§2), including real-world deployment scenarios and key technical components. We then present the motivation and threat model of AFR tools (§3), and our framework for analyzing existing AFR tools (§4). We then discuss existing AFR proposals targeting each stage, i.e., *data collection* (§5), *data processing* (§6), *feature extractor training* (§7), *identity creation* (§8), and *query matching* (§9). Finally, we identify desirable properties of effective AFR systems, and map them to points in the design space (§10). Finally, we discuss challenges and directions for AFR research (§11).

Unresolved Ethical Questions: The broad deployment of facial recognition systems (and by extension, AFR systems) is fraught with ethical challenges, not the least of which are significant biases against women and people of color [36]. While we discuss ethical tensions surrounding AFR systems in §11.2, we do not make assertions about how (and whether) AFR tools should be used. Development and adoption of AFR tools are driven by backlash against biased and misused facial recognition systems. Though their legal and ethical implications are yet-unknown, we believe that AFR tools are here to stay. Consequently, an analysis of their strengths and limitations is crucial to advancing the ongoing debate about their use and the place of facial recognition in our world.

2. Facial Recognition: Terminology, Design Stages and Deployment

To provide context for later discussions, we now give a high-level overview of today's facial recognition (FR) systems and their real-world implementations. Our goal is to describe modern FR systems targeted by today's AFR systems, their key operational stages, and how these FR systems are being deployed around the globe. Together, these provide a framework that we will use for analyzing AFR systems later in §4, by examining critical points of direct interaction between users and FR systems.

2.1. Modern FR Systems

FR systems identify people by their facial characteristics, generally by comparing an unknown face in an image (or a video) against a database of known faces. The technology has evolved significantly over the past two decades, resulting in many design variants [37]. Today, the state-of-the-art and widely adopted FR systems employ deep neural networks (DNNs) to extract unique features from a given face. Since existing AFR systems mainly target these modern FR systems, this work focuses on DNN-based FR/AFR systems. The main differences between older and newer FR methods lie in (1) the feature extraction methods (e.g. statistical methods like PCA or LDA [38, 39] vs. DNN-based feature extractors) and (2) scale However, the fundamental FR stages remain the same in both older and newer FR systems – face images must still be collected, processed, and recognized. Consequently, the framework laid out in this paper could be applied to older FR/AFR systems if desired.

Terminology. In this paper, we represent a modern FR system as $\mathbb{F} = \{\mathcal{G}, \mathcal{F}, \mathcal{C}, \mathbf{D}\}$, whose goal is to associate a query image x_I with its true identity I. Specifically,

- *Query image* (x_I): a face image to be identified by $\mathbb{F}$.

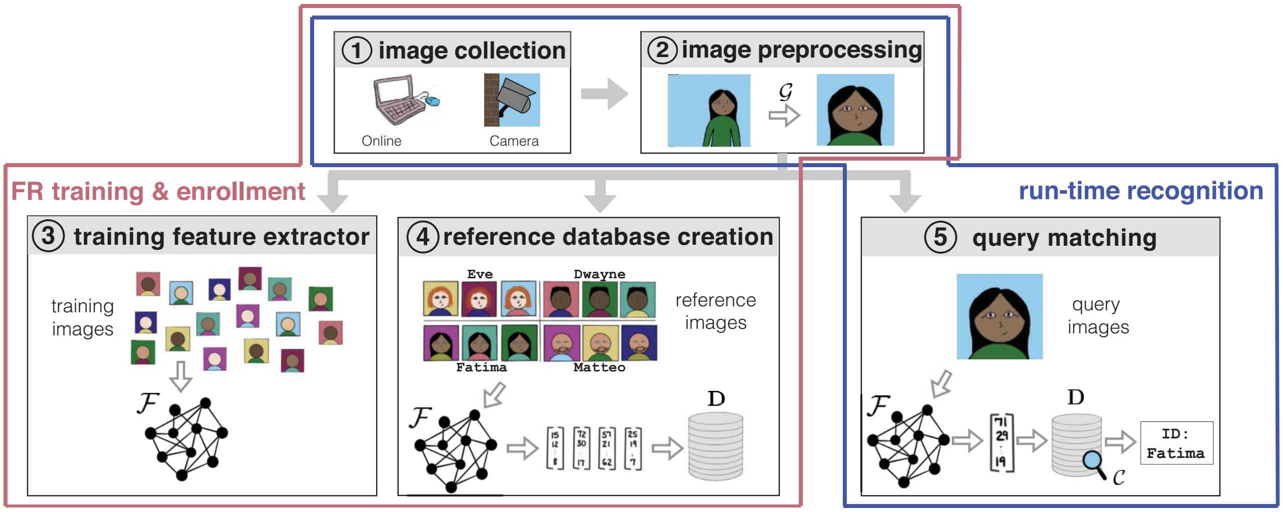

Figure 2. We propose to divide the operational pipeline of a FR system $\mathbb{F}=\{\mathcal{G},\mathcal{F},\mathcal{C},\mathbf{D}\}$ into a set of five operational stages ① to ⑤. They encompass the five critical points of direction interaction between users and FR systems. Later we will use this framework for analyzing AFR systems.

- *Preprocessing engine* ($\mathcal{G}$): a processing function that prepares raw face images for the FR task, e.g., detecting and cropping out individual faces.
- *Face feature extractor* ($\mathcal{F}$): a DNN that converts a face image into a *feature vector*, a mathematical representation of the person's unique facial features.
- *Reference database* ($\mathbf{D}$): a collection of face images and their feature vectors of known identities, e.g., x_I^R (ground truth images of user I) and $\mathcal{F}(\mathcal{G}(x_I^R)) = v_I$.
- *Run-time face classifier* ($\mathcal{C}$): this function runs a query search to match the query image $\mathcal{F}(\mathcal{G}(x_I))$ against $\mathbf{D}$. If the closest feature vector v_I is sufficiently similar, then the query image is identified as I. Ideally, it should produce $\mathcal{C}(\mathcal{F}(\mathcal{G}(x_I)),\mathbf{D}) = \mathcal{C}(\mathcal{F}(\mathcal{G}(x_I^R)),\mathbf{D})$, where x_I is a previously unidentified image of I and x_I^R is a ground-truth reference image of I.

It should be noted that the terminology used to describe a FR system can vary across the literature. We list some alternative terms in Figure 1. The terms we choose to use in this paper are, we believe, most familiar to the security community.

Face recognition vs. face verification. Here, we note the distinction between *face recognition* and *face verification*. Face verification is widely used to authenticate users on mobile devices (e.g., FaceID on iPhones) by comparing a user's face feature to the stored face feature of the authorized user. While the two systems apply similar techniques to analyze face images, facial verification systems require *user consent* for deployment while many FR systems operate without user consent. As such, most AFR systems target FR systems rather than face verification systems. *Furthermore, face verification systems can be viewed as a special case of FR systems, where the reference database only has a single user.* Therefore, in this paper we do not explicitly consider facial verification or its disruption.

2.2. Breaking FR into Operational Stages

We now examine the FR operational pipeline and divide it into a set of *operational stages* to help frame our discussion of AFR tools. These operational stages correspond to specific subtasks in FR, which encompass *the five critical points of direct interaction between users and FR systems.* Figure 2 depicts the five operational stages of a FR system $\mathbb{F} = \{\mathcal{G},\mathcal{F},\mathcal{C},\mathbf{D}\}$. We discuss each stage below and revisit them as a framework to analyze AFR tools in §4.

The overall operation of a FR system includes two phases: a *training & enrollment* phase where the system builds (or acquires) a face feature extractor and creates a reference database of known identities, and a *run-time recognition* phase where the FR system identifies an unknown face. As we show below, the training & enrollment phase employs stage ①–④, while the recognition phase employs stage ①, ② and ⑤.

Stage ①: collecting face images. Face images primarily come from two sources: online image scraping [40] or physically taking a photo of a person [1, 8]. We discuss sources of face images for FR systems in further detail in §2.3.

Stage ②: preprocessing raw face images via $\mathcal{G}$. Raw images obtained from stage ① are often poorly structured (e.g., varying face sizes, bystanders in background). To make downstream tasks easier, $\mathbb{F}$ employs an image preprocessing engine $\mathcal{G}$ that uses face detection (e.g., automated face cropper [41]) to remove background and extract each individual face, followed by a data normalization process [42, 43, 44].

Stage ③: training a feature extractor $\mathcal{F}$. The crucial element of DNN-based FR systems is the feature extractor $\mathcal{F}$ 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 [44] in 2015: adding an *extra* loss function during $\mathcal{F}$ training to directly optimize for large separations between different faces in the feature space. Followup works explore alternative loss functions and architectures to further improve the accuracy of FR systems (e.g., [42, 43, 45]).

To maximize efficacy, $\mathcal{F}$ 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 [46] and government agencies [47, 48], opt to purchase or license a well-trained feature extractor (e.g. [49, 50, 51, 52, 53, 54, 55, 56]). We refer to images used in stage ③ as *training images*.

Stage ④: creating a reference database $\mathbf{D}$. FR systems need a large database of known (labeled) faces in order to identify unknown (unlabeled) faces. As a result, FR operators must build a reference database $\mathbf{D}$ of people they want to recognize, by first collecting and preprocessing labeled face images of these individuals, and then passing them to $\mathcal{F}$ to obtain feature vectors. The reference database holds the (feature vector, identity) pairs [40, 57, 58]. We refer to images used in stage ④ as *reference images*.

Stage ⑤: recognize the face in a query image via $\mathcal{C}$. At run-time, the FR system takes in and preprocesses (via $\mathcal{G}$) a query image (i.e., an unidentified face image), extracts its feature vector via $\mathcal{F}$, and queries the reference database $\mathbf{D}$ to locate a match (if any). If the feature space distance (e.g., L_2 or cosine) of the query image is sufficiently close to a stored entry in $\mathbf{D}$, the system outputs a match. In this paper, we represent this process by the classifier $\mathcal{C}$.

2.3. FR Deployment and Data collection

In recent years, entities across the globe have adopted and deployed FR systems for various applications. This wide adoption was triggered by significant accuracy improvements of FR systems, largely due to new training methods [44] and more powerful neural network architectures [59]. Deployment scenarios of these FR systems, along their data sources, have informed AFR tool development. Thus, to contextualize AFR proposals, we briefly examine how FR is used in the real world and from where its images (i.e., training/reference/query images) are drawn.

Deployment scenarios. Both public and private entities use FR for a variety of purposes. We list some examples in Table 3. Public (e.g., government-based) FR use cases range from criminal identification[2][60], civilian surveillance [2, 61] and border control [62], to video game use tracking [63] and COVID lockdown enforcement [64]. For a broader exploration of government uses of facial recognition, we refer the reader to [65]. Private entities have also integrated FR into their security and commerce pipelines. The most common private FR use cases are enhancing store or office security, but other examples abound (see Table 3).

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 today.

Training images (used to train the feature extractor $\mathcal{F}$) often come from a mix of academic training datasets (e.g., [66, 67, 68, 69]), proprietary data, and public data scraped from social media accounts, according to a report of the US Government Accountability Office [47].

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). Table 4 shows a list of known reference image sources for some well-known FR operators.

Query images, or faces to be identified by the FR system at run-time, can come from both online and physical sources. Some known sources include social media, police body cams, mug shots, corporate surveillance systems, state ID images, and passport photos [48]. After identification, query images are sometimes fed back into the reference database, either to enhance existing feature vectors or to 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 [70]. Similar techniques are used by several Chinese companies [10, 71].

3. 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.

3.1. 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 [5] without user knowledge or consent. *Second*, increased government use of FR systems has caught the attention of citizens who have raised significant concerns about the long-term effects of FR on privacy and freedom of expression [15, 72]. *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, 73, 74].

Consequently, public sentiment about FR is mixed and, especially in western countries, trending negative [75, 76, 77, 78]. This shift in public opinion, combined with the forces noted above, has motivated researchers to create various AFR tools to counteract unwanted FR systems.

3.2. Threat Model of AFR

AFR tools are used by a person P to combat a FR system $\mathbb{F} = \{\mathcal{G}, \mathcal{F}, \mathcal{C}, \mathbf{D}\}$. In this context, P takes the role of an attacker and acts against $\mathbb{F}$. P's goal is to prevent recognition

[2] Recently, police departments around the US have drawn fire for their use of highly unregulated FR software like Clearview.ai [50].

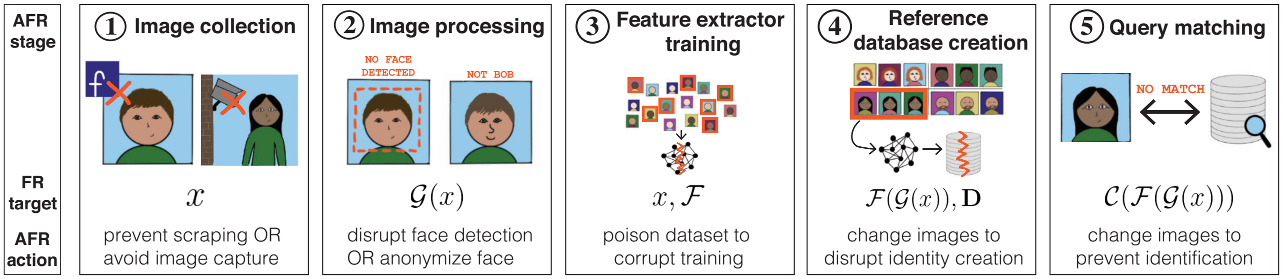
AFR stage
FR target
AFR action
① Image collection
x
prevent scraping OR avoid image capture
② Image processing
NO FACE DETECTED
NOT BOB
$\mathcal{G}(x)$
disrupt face detection OR anonymize face
③ Feature extractor training
$x, \mathcal{F}$
poison dataset to corrupt training
④ Reference database creation
$\mathcal{F}(\mathcal{G}(x)), \mathbf{D}$
change images to disrupt identity creation
⑤ Query matching
NO MATCH
$\mathcal{C}(\mathcal{F}(\mathcal{G}(x)))$
change images to prevent identification

Figure 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 §2.2 and present AFR strategies per stage by the attack target, action, and desired effect.

by $\mathbb{F}$, i.e., given an image x_P of P, a successful AFR tool should cause $\mathbb{F}$ to produce $C(F(G(x_P))) \neq P$.

Proposed AFR tools generally make the following assumptions about each party:

- P has no special access to or authority over $\mathbb{F}$, but wishes to evade unwanted identification by modifying or otherwise controlling their own face images.
- P wishes to avoid facial recognition, but also may wish for their images to remain useful for other purposes. For example, if P posts a headshot on a personal website, they would like to ensure that the headshot not be scraped and used in a FR engine but also that their face remains recognizable to website visitors. Thus, P prefers AFR tools which maximize AFR protection while minimizing image disruption.
- $\mathbb{F}$'s goal is to either create or maintain an accurate facial recognition operation. Furthermore, $\mathbb{F}$ operates *at scale* and does not specifically target P for identification.

Implications. We also explore the real-world implications of the above threat model.

(1) *Assuming AFR tools operate on images* – Our study focuses exclusively on image-based AFR tools that a user P can deploy on their own. These image-based designs, which operate either directly on images or on systems that collect/process images, dominate the current set of AFR proposals. On the other hand, a user P may, depending on the context, be able to use other means (e.g., legal action) to fight unwanted facial recognition.

(2) *Assuming $\mathbb{F}$ does not specifically target P for recognition* – Existing AFR tools are designed to fight large-scale FR systems. This is because, from a practical standpoint, if system $\mathbb{F}$ wishes to specifically recognize a user P, there are much more efficient options than using a general, large-scale FR system. Therefore, most current AFR tools are not designed to withstand this level of scrutiny. If $\mathbb{F}$ makes a more targeted effort to identify P, such as hiring a private investigator, current AFR tools will likely fail.

(3) *Assuming AFR tools minimize perturbations.* This study focuses on AFR tools which introduce *minimal perturbations* to images (as measured by L_P norms). This decision is grounded in prior work showing that users are more likely to use privacy-preserving tools with minimal overhead [27, 79, 80]. AFR tools which do not seek to minimize perturbations are not addressed in this work. This is a limitation of our work, and future work should consider AFR tools which use metrics beyond L_P norms to measure image distortion.

4. A Stage-Based AFR Framework

We now discuss and analyze existing AFR proposals. To do so, we propose and use a *stage-based* framework to categorize AFR strategies. As discussed in §2, a FR system $\mathbb{F} = \{\mathcal{G}, \mathcal{F}, \mathcal{C}, \mathbf{D}\}$ operates in 5 distinct stages that correspond real-world actions (e.g. image capture, pre-processing, feature extraction, etc.). In each stage, $\mathbb{F}$ interfaces with P and the broader world as it collects, processes, and uses image data. Such interfaces represent possible points at which P can act against $\mathbb{F}$. Specifically, an AFR tool can attack the component(s) of $\mathbb{F}$ relevant to any of the 5 stages, including images x, preprocessor $\mathcal{G}$, feature extractor $\mathcal{F}$, reference database $\mathbf{D}$, and classifier $\mathcal{C}$. With this in mind, Figure 3 demonstrates the attack actions and goals when AFR tools target each FR stage.

4.1. AFR Strategies per Stage

Since the five FR stages ①–⑤ encompass the points of direct interaction between P and $\mathbb{F}$, 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.

Attacking ①. In the image collection stage, labeled and/or unlabeled images x are collected for use by $\mathbb{F}$, either by physically taking photos or scraping online images. Labelled images can be used as training or reference images to build a FR system, while unlabelled images can be used as query images. When targeting this stage, AFR tools focus on disrupting the data collection process to prevent $\mathbb{F}$ from acquiring usable face images x_P of P.

Attacking ②. In the second stage, $\mathbb{F}$ uses $\mathcal{G}$ to preprocess collected face images using a series of digital transformations, e.g., face detection, background cropping, and normalization. AFR tools deployed at this stage target $\mathcal{G}$ to render the processed images unusable, either by breaking the preprocessing functions (e.g., preventing faces from being detected) by injecting noise and artifacts onto the images or removing P's identity information from the images. We denote these different actions as (2a) and (2b), respectively.

AFR system	**Year released**	**Stage targeted**	**Attack scenario**				
			P's knowledge of $\mathbb{F}$	*P's operating context*	*Targeted/ Untargeted*	*Tested on real-world FR*	Unique Property
Anti-scraping [98-102]	2021	①	-	Digital	UT	-	Prevent large-scale image scraping
Data Leverage [81]	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. [82]	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
Hu et al. [83]	2021	②a	WB	Digital	UT	-	Adversarial patch on object detectors
Treu et al. [25]	2021	②a	BB	Digital	UT	-	Adversarial clothing on face detectors
DeepPrivacy [84]	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 [85]	2021	②b	BB	Digital	UT	-	GAN-based face blurring (imperceptible)
Evtimov et al. [86]	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
Fu et al. [87]	2021	③	BB	Digital	UT	-	Data poison by unlearnable data
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. [88]	2013	⑤	BB	Physical	UT	-	Make-up
Sharif et al. [89]	2016	⑤	Both	Both	Both	F++	Adversarial patch on wearable accessories
Dabouei et al. [90]	2018	⑤	WB	Digital	UT	-	Adversarial attack distorts face landmarks.
Zhou et al. [91]	2018	⑤	WB	Physical	Both	-	Projected adversarial IR patterns
Dong et al. [92]	2019	⑤	BB	Digital	T	TN	Black-box adversarial perturbation.
Zhu et al. [93]	2019	⑤	Both	Digital	Both	-	Adds eye makeup with GAN.
AdvHat [30]	2019	⑤	WB	Physical	UT	-	Printed sticker on hat.
AdvFaces [94]	2019	⑤	BB	Digital	Both	-	GAN-based adversarial attack.
VLA [95]	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. [96]	2021	⑤	WB	Digital	UT	-	Brightness-agnostic adversarial perturbations
Yang et al [97]	2021	⑤	BB	Digital	UT	TN	Corrupts features of faces

TABLE 1. 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 ③. Since stage ③ is dedicated to training $\mathbb{F}$'s feature extractor $\mathcal{F}$, AFR tools targeting this stage seek to degrade the accuracy of $\mathcal{F}$ by poisoning its training images.

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

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

4.2. Taxonomy of Existing AFR Proposals

Using our stage-based analysis framework, we now present a comprehensive taxonomy of existing AFR proposals in Table 1. 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 P's knowledge of $\mathbb{F}$ (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 been tested against real-world FR systems, and any unique or notable features of the AFR tool. We note that the majority of AFR tool users may not care for or need a *targeted* AFR misclassification result, but we include targeted attacks for completeness, as they represent the most user-controlled version of an AFR tool.

There is 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 research. We also notice that 7 out of 30 proposals assume a "white-box" access to $\mathbb{F}$'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 1 serves as a comprehensive summary of current AFR proposals, which we will refer to throughout the paper.

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

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

Adversarial ML and AFR. A significant portion of AFR tools listed in Table 1, e.g., those targeting (2a) and ③ – ⑤, apply adversarial machine learning (AML) techniques like poisoning or evasion attacks to thwart $\mathbb{F}$. Consequently, a significant portion of this paper is devoted to discussing the pros and cons of AML-based approaches to AFR. On the other hand, since AFR tools targeting stage ① and (2b) are inherently non-AML based, our analysis is not limited to only AML-enabled AFR tools. Since each of the five FR stages represents a viable attack vector for AFR tools, our analysis covers AFR tools targeting any stage.

4.3. Roadmap of Our Analysis

Using the stage-based framework, we conduct a detailed analysis of existing AFR proposals. First, we discuss in greater detail how existing AFR proposals attack each of the five stages (§5 – §9). In each section, we describe the goals of $\mathbb{F}$ and P, the challenges of targeting this particular stage, the existing proposals, and a summary of the progress made in this direction. Next, we conduct a meta-analysis of AFR strategies across the five stages (§10), and discuss what we see as the major technical and broader social/ethical challenges facing future AFR development (§11).

5. Attacking ① to Disrupt Data Collection

We start by examining AFR methods that allow P to attack $\mathbb{F}$ by disrupting the process of face data collection.

Goals and Challenges. In this data collection stage, $\mathbb{F}$'s goal is to obtain usable face images x from online or physical sources. Often, $\mathbb{F}$ aims to collect high quality images of millions or billions of people (e.g., Clearview.ai [5]). $\mathbb{F}$ uses labeled images to build the reference database and/or train the feature extractor. By using AFR tools, P's goal is to prevent their face images x_P from being collected for use by $\mathbb{F}$, either as (labeled) reference images or (unlabeled) query images. They can apply online or physical evasion/disruption techniques to do so.

The key challenges facing AFR proposals targeting this stage are that (1) they need to be aware of and adapt to $\mathbb{F}$ who continues to innovate techniques for data collection; and (2) cameras are widely deployed in the real-world, making it challenging to avoid image capture by cameras.

5.1. Current Solutions

Face images can come from two sources: scraping online images or physically capturing faces using cameras. Image scraping refers to the collection of images posted online that were captured by someone who is not the data collector (e.g. camera is operated by the subject, subject's friend, etc). Picture-taking refers to images taken directly by the data collector. Thus we divide AFR tools acting at this stage into two categories: preventing scraping and preventing capture.

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

- *Anti-scraping by online platforms.* Anti-scraping techniques have been widely studied in the security community [98, 99, 100, 101, 102]. Techniques such as rate limits, data limits, ML-based scraping detection are already used by online platforms [103]. However, a significant portion of scraping still goes undetected as scrapers develop more sophisticated tools to bypass detection [103].
- *Data leverage by users.* P could try to prevent $\mathbb{F}$ 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 [32, 81, 104]. While not specifically aimed at facial recognition, these proposals offer alternative models for online engagement while protecting user data.

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 [105]. Furthermore, face photos taken by on-street surveillance cameras are increasingly used by commercial or government facial recognition systems [1, 9, 106, 107, 108], 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 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 [109]; 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 [110].
- *Camera disruption.* Without physically breaking cameras, human users can prevent cameras from capturing (usable) images by simply shining laser lights at them [72]. Other methods include covering cameras with fabric or stickers.

5.2. Discussion of Stage ① Solutions

Privacy/Utility Trade-offs. Evading data collection requires both fine-grained control over one's online identity and awareness of when/how pictures are being taken, making it difficult for users to deploy these tools without significantly limiting either their online or physical activities. Furthermore, there are cases where evading data collection is simply impossible, i.e. mandatory pictures posted on an employer's website. Anti-scraping tools can also decrease the utility of the service provider, as such tools can have false positives and a high deployment cost.

Summary of Progress. Existing AFR proposals against stage ① make headway in addressing key challenges. Adaptive anti-scraping techniques like [103] definitely raise the bar for attackers. Furthermore, anti-data collection methods like [110] have shown that it is possible, with careful action, to evade image capture even in robust surveillance systems. Future AFR development against stage ① can seek to improve both data controls and camera awareness by individual users.

6. Attacking ② to Disrupt Face Pre-processing

In stage ②, $\mathbb{F}$ processes raw face images with $\mathcal{G}$ to facilitate further operations in stages ③, ④, and ⑤. AFR proposals targeting this stage seek to disrupt $\mathcal{G}$ by transforming an image x into x' such that processed face images $\mathcal{G}(x')$ are "unusable" by subsequent FR stages.

Goals and Challenges. $\mathbb{F}$'s goal is to use $\mathcal{G}$ to obtain well-structured face images from many raw images. P's goal is to either prevent their face being detected/extracted from raw images by $\mathcal{G}$ or to anonymize their face in these images.

The main challenge for AFR proposals targeting stage ② is how to achieve *anonymization without distortion*. That is, when modifying P's images to either evade $\mathcal{G}$'s face detection or to remove identity information, the modified images should still resemble P to remain *useful* to P. An additional challenge is overcoming (adaptive) defenses deployed by $\mathbb{F}$ to protect $\mathcal{G}$.

6.1. Current Solutions

Preventing Face Detection. Face detection extracts well-centered headshots from raw images. The commonly used face detection systems [41] rely on DNNs to accurately infer the location of faces in an image. To prevent effective face detection/extraction, the AFR goal is to produce an adversarial x' such that $\mathcal{G}(x') = z$, where z is a useless result that cannot be passed on to $\mathcal{F}$. To create x', existing AFR tools leverage "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., [111, 112, 113, 114]). 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 P's knowledge of $\mathbb{F}$ (e.g. white vs. black box, see Table 1).

AFR tools using adversarial perturbations can be subdivided based on how the perturbation is added to images. They can be directly added to digital images if P has direct access to these images or fabricated as physical objects that P can wear (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., [25, 28, 83]). 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, 109], t-shirts [24, 26], 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 [82].

Anonymizing Faces 2b. P can also *anonymize* their face images to remove identity information. In this setting, P creates an x' such $\mathcal{G}(x') \neq \mathcal{G}(x)$, i.e. the result after processing is still usable but represents a fake identity. *Physical anonymization* can be easily achieved by wearing masks, hats, makeup, etc, which overlaps with "avoiding image capture" in ① discussed in §5. Leaving proposals for *digital anonymization* use generative adversarial networks (GANs) [115] and differential privacy [116]. Several such proposals use GANs to transform face images into latent space vectors, modify those vectors to remove identity information, and reconstruct the images from the modified vectors [17, 84, 85]. The modified faces still look human but are anonymized to prevent accurate identification. Another proposal, IdentityDP [18], uses similar techniques but also claim to provide differentially private identity protection.

6.2. Discussion of Stage ② Solutions

Privacy/Utility Trade-offs. Many stage 2a proposals address the "usability" challenge by formatting adversarial patches against $\mathcal{G}$ as wearable clothing/objects. However, wearing this special clothing, which can appear bizarre, may not be desirable for the average person. Current proposals against 2b provide anonymity but tend to produce anonymized faces that do not resemble the original face, with significantly altered shape, skin tone, hair color, etc. These images lack many functionalities of traditional images, e.g. image sharing, preserving memories, etc.

Summary of Progress. Many AFR proposals targeting 2a have been tested against real-world object detectors like YOLOv2, demonstrating their real-world efficacy. However, several defenses against these patches have emerged recently (e.g., [117, 118, 119]), although only one [117] has been tested against physical adversarial patches like the ones used by AFR tools [24, 26]. Further work is needed to determine if AFR proposals against 2a can resist these defenses that $\mathbb{F}$ can use to protect $\mathcal{G}$.

7. Attacking ③ to Corrupt Feature Extractor

All FR systems require an effective feature extractor $\mathcal{F}$ to distinguish between faces. AFR proposals attacking stage

③ focus on corrupting the training of $\mathcal{F}$ to produce an unusable extractor $\mathcal{F}'$.

Goals and Challenges. Here, $\mathbb{F}$'s goal is to train a high-quality feature extractor $\mathcal{F}$ using available training data, so that faces can be accurately identified by their feature vectors extracted by $\mathcal{F}$. Thus P's goal is to prevent $\mathbb{F}$ from training an effective $\mathcal{F}$ by corrupting the training data.

There are two key challenges facing AFR proposal targeting stage ③. The first is minimizing the distortion to training face images introduced by the corruption process while maintaining the corruption efficacy. The second is corrupting $\mathcal{F}$'s training without requiring full dataset control.

7.1. Current Solutions

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 [120, 121, 122, 123, 124]. 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, recent works [20, 87] render data "unlearnable" by a DNN model. This noise misleads the model into thinking that data have already been learned, thwarting necessary parameter updates. When a user submits their "unlearnable" face images as a training image for the $\mathcal{F}$, the extractor will not learn anything to improve its performance. Training an effective $\mathcal{F}$ requires millions or even billions of face images [42, 43, 44], and with a sufficient number of unlearning training examples, $\mathcal{F}$ will not meet the accuracy level required for practical deployment.

Adding adversarial shortcuts. A related proposal from Evtimov *et al.* [86] 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.

7.2. Discussion of Stage ③ Solutions

Privacy/Utility Trade-offs. The biggest utility drawback of stage ③ proposals is that they require significant effort to corrupt the training dataset. Most proposals require that P control much of the training data to render $\mathcal{F}$ unusable. Such high levels of control would prohibit individual users from using these AFR tools. In addition, $\mathbb{F}$ can discard a corrupted dataset and use other data sources to train their model, once they discover the presence of corruption.

Summary of Progress. Despite utility challenges, existing proposals have shown that, with a sufficient level of dataset control, it is possible to render $\mathcal{F}$ unusable by adding minimally visible perturbations onto training face images. For example, AFR tools based on adversarial shortcuts [86] are effective when they can corrupt the entire training dataset. Others [20] can reduce $\mathcal{F}$'s accuracy on specific classes in a FR model by 16%.

8. Attacking ④ to Corrupt Database

In stage ④, $\mathbb{F}$ uses $\mathcal{F}$ to create a reference database $\mathbf{D}$ of labeled face feature vectors that will facilitate identification of unidentified faces. AFR tools targeting this stage seek to fill $\mathbf{D}$ with incorrect face/label mappings, so that $\mathbb{F}$'s classifier $\mathcal{C}$ cannot identify P's query images as P. Thus, when a true image x_P is presented to $\mathbb{F}$'s system for identification, the corrupted database $\mathbf{D}'$ produces $\mathcal{C}(\mathcal{F}(\mathcal{G}(x_P)), \mathbf{D}') = I$, where I is an incorrect identity, $I \neq P$.

Goals and Challenges. In this stage, $\mathbb{F}$'s goal is to create a reference database containing accurate copies of feature vectors (produced by $\mathcal{F}$) of people $\mathbb{F}$ wishes to recognize. P's goal is to prevent $\mathbb{F}$'s feature extractor $\mathcal{F}$ from creating an accurate feature vector which $\mathcal{C}$ can match to query images of P. Note that this can also be achieved by corrupting the training data/process of $\mathcal{F}$ in stage ③, as discussed in §7. Here, we differentiate from §7 by assuming that $\mathcal{F}$ is a well-trained feature extractor. Thus, P attacks $\mathbb{F}$ by modifying/manipulating the reference images of P that $\mathbb{F}$ uses to create its reference database.

Stage ④ based AFR tools must first address the base case challenge of modifying P's reference images to produce incorrect feature vectors while minimizing the distortion of those images. They also face two advanced challenges. First, they must maintain high performance when $\mathbb{F}$ has some original, unmodified face images of P already enrolled in its database $\mathbf{D}$. Second, protection must persist when P makes incorrect assumptions about $\mathbb{F}$'s system, especially its extractor $\mathcal{F}$ and classifier $\mathcal{C}$, or when $\mathbb{F}$ adapts.

8.1. Current Solutions

Existing AFR proposals in this category focus on poisoning feature vectors before they are stored in $\mathbf{D}$. The specific poisoning techniques depend on the underlying assumptions about $\mathbb{F}$'s classifier $\mathcal{C}$.

Assuming $\mathcal{C}$ uses classification-based matching. A recent AFR proposal, Fawkes [19], assumes $\mathcal{C}$ is a shallow classification layer added to $\mathcal{F}$. Fawkes seeks to corrupt the final classification output by "cloaking" (or poisoning) reference images of P, i.e. shifting their feature vectors away from the correct representation by adding imperceptible perturbations to P's reference images [19]. When $\mathcal{C}$ is trained on these shifted feature vectors, $\mathbb{F}$ will learn to associate incorrect feature spaces with P's identity, producing wrong matches for P's (uncloaked) query images at run-time.

Assuming $\mathcal{C}$ uses nearest neighbor-based matching. Two other AFR proposals, LowKey [22] and FoggySight [21], assume $\mathcal{C}$ is a K-nearest neighbors algorithm. LowKey [22] adds digital adversarial perturbations to change the feature representation of P's reference images (similar to Fawkes). These perturbed images create a reference feature vector for P that is different from those of P's run-time query images, thus thwarting $\mathcal{C}$. 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.

8.2. Discussion of Stage ④ Solutions

Privacy/Utility Trade-offs. Most Stage ④ proposals add perturbations directly to images. Several proposals discuss how stronger (more visible) perturbations yield stronger AFR protection (i.e. [19, 22]). Visible perturbations may lower the utility of protected images, especially if they are meant to be posted on social media sites. More advanced optimization techniques may help reduce perturbation size at stronger protection levels, but this visibility/protection trade-off seems inevitable.

Summary of Progress. It is encouraging that all proposals listed in the above analysis have demonstrated success on the key task of corrupting feature vectors of P, i.e., the base case. Overcoming the two advanced challenges discussed remains an area for future work. Some proposals provide limited protection when $\mathbb{F}$ has obtained original, unmodified feature vectors of P (e.g. [19]), but not all proposals have considered this possibility. Second, all existing proposals assume knowledge of $\mathcal{C}$ and/or $\mathcal{F}$, necessitating further work to determine how/if incorrect assumptions of $\mathbb{F}$ would affect AFR performance. A final challenge is evaluating the long-term robustness of stage ④-based AFR mechanisms against an adaptive $\mathbb{F}$. Recent work (discussed in §11) suggests that a continuously adapting $\mathbb{F}$ may always (or eventually) "win" against AFRs targeting ④.

9. Attacking ⑤ to Evade Identification

The final set of AFR tools aims to prevent run-time query image identification by $\mathbb{F}$'s classifier $\mathcal{C}$, by producing distorted images x'_P that mislead the final classifier outcome, e.g. $\mathcal{C}(\mathcal{F}(\mathcal{G}(x'_P)), \mathbf{D}) = I \neq P$. These methods can provide one-time protection for users who believe their images are already enrolled in $\mathbf{D}$. 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 §8). However, current AFR proposals targeting this stage focus strictly on evasion and do not consider this joint possibility.

Goals and Challenges. In this run-time reference stage, $\mathbb{F}$**'s goal** is to use $\mathcal{C}$ to identify the face in a query image. P**'s goal** is to alter their query image so $\mathcal{C}$ cannot match it to their corresponding feature vector in $\mathbf{D}$. We assume $\mathbb{F}$'s reference database $\mathbf{D}$ contains accurate feature vectors of P.

There are two key challenges for stage ⑤-based AFR proposals. The first is achieving successful evasion without significant image distortion. Additionally, proposals must overcome defenses deployed by $\mathbb{F}$ to protect $\mathcal{F}$ and/or $\mathcal{C}$.

9.1. Current Solutions

Adversarial perturbations have been the dominant method for evading DNNs 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 P wear them as physical objects. While these methods echo those described in §6, they focus on thwarting recognition/classification rather than face detection. Earlier proposals [88, 89] use adversarial makeup and eyeglasses to cause incorrect classification by $\mathcal{C}$. 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 [29, 91, 95].

Digital evasion techniques. Here P digitally modifies their unlabeled (online) face images to prevent them from being accurately classified by $\mathcal{C}$. Most proposals in this category apply traditional adversarial perturbation generation techniques to create minimally visible perturbations that cause $\mathbb{F}$'s feature extractor to produce misleading feature vectors. Their generation process varies depending on assumptions of $\mathcal{C}$'s behavior: a shallow classification layer vs. nearest neighbor based matching [90, 92, 93, 96].

More recent proposals are designed to be more robust to real-world FR systems (i.e. joint optimization on multiple feature extractors, etc.) [23, 27, 97]. Another recent proposal [94] uses a GAN to generate adversarial perturbations rather than using optimization techniques.

9.2. Discussion of Stage ⑤ Solutions

Privacy/Utility Trade-offs. For physical evasion techniques, a significant usability challenge comes from the possibility of real-time recognition. In order to ensure physical evasion tools are effective, a user must wear them in all circumstances where cameras might be present. As with Stage ④, there also exists here a trade-off between perturbation size and evasion success for digital evasion techniques. Reducing the amount of perturbation needed to evade recognition remains an active area of research.

Summary of Progress. So far, existing works have focused on addressing the first key challenge, and have evasion success with minimally visible perturbations on query images. Addressing the second challenge, or understanding how AFR tools interact with existing defenses against evasion attacks, remains an open area of research. Defenses against evasion attacks like the ones listed above are being released regularly (e.g. [111, 125, 126, 127]), only to be broken by new attacks (e.g. [128, 129, 130]). No defenses have yet been explicitly proposed for these attacks, but the general trend suggests this may be possible.

10. Goals and Tradeoffs in AFR Design

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 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 bene-*

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

TABLE 2. EVALUATING AFR TOOLS USING FIVE PROPERTIES, WHERE THE TOOLS ARE GROUPED BY THE FR STAGE THEY TARGET.

fits and limitations of AFR tools that target each specific framework stage?" Or, an alternative form of the question might 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. While we consider each stage in isolation, it might be possible for an AFR tool to target multiple stages, gaining a combination of benefits (and limitations).

10.1. Five AFR Design Properties

When considering design properties of AFR tools, we assume that efficacy is a given. 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 others?
4) *Minimal friction for user P*, minimizing cost for P 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

10.2. Implications of Properties for AFR Design

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 2 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 P from unwanted FR both initially and as FR evolves.

●: 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, P can achieve long-term robustness by consistently undermining the face data pipeline of $\mathbb{F}$. AFR① and AFR② can both prevent any face image of P to be included into $\mathbb{F}$'s pipeline. AFR④ can corrupt $\mathbb{F}$'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* 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 $\mathbb{F}$ 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 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 $\mathbb{F}$ has some ground truth information about P's facial features. However, the addition of protected images to the reference database can slowly disguise P's true features, 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 P against $\mathbb{F}$ who has already obtained and processed unprotected images of P.

Property 3: No reliance on 3rd party to operate. Ideally, an AFR tool can be operated by a user 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 help 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 data to a 3rd party.

●: AFR②, AFR④, AFR⑤ AFR tools in these three groups all rely on adding certain perturbations on face images, which P can do without outside assistance.

?: AFR①, AFR③ For those AFR① seeking to prevent online data scraping, they rely on the assistance of image sharing platforms. Similarly, disrupting the training $\mathcal{F}$ requires coordinated effort across many users, since P alone contributes relatively few images to the training data.

Property 4: Minimal friction for P. This usability-related property measures what 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 [79, 80].

◐: AFR①, AFR②, AFR③, AFR④, AFR⑤ So far, existing AFR tools all introduce some level of "disruption" to P, whether by adding visual noise, perturbations or transformations to P's online photos that distorts them, requiring P to always wear odd makeup/clothes/accessories, or necessitating more powerful computing hardware/services to implement the AFR tool against continually evolving $\mathbb{F}$. 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 P's AFP protection would affect other users. Intuitively, P can protect themselves by forcing $\mathbb{F}$ to fail (give a null or uninformative result), or by intentionally tricking $\mathbb{F}$ to recognize them as another person P'. Depending on the context, the latter may negatively affect P', producing potential social risks (see §11.2 for detailed discussions on social challenges of AFR).

●: AFR①, AFR② These AFR tools disrupt the data pipeline of $\mathbb{F}$, and thus, have no impact on other users.

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

11. 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 §10, 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 roadmap.

11.1. Technical Challenges

TC 1: Reliance on AML-based tools. The majority of AFR proposals, especially those targeting stages ② – ⑤, employ techniques from adversarial machine learning (AML), which have several key limitations. First, while AML tools have exhibit high performance, they have not provided provable guarantees of protection. Second (and related), AML-based protections can be defeated by adaptive FR systems. For example, $\mathbb{F}$ could adversarially train the feature extractor $\mathcal{F}$ [131, 132] to be more robust against adversarial examples, thus defeating AFR tools against stages ③ or ④. $\mathbb{F}$ could also remove adversarial perturbations from face images before processing them or adding them to the reference database [133], circumventing AFR tools that target stages ② or ⑤.

Potential Directions. More advanced perturbation generation methods may help increase short-term efficacy of AML-based 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 alternative techniques that provide 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 DNN (e.g. non-AML AFR tools). The second is to switch from "misleading" $\mathbb{F}$ with "minor" image modifications to completely disabling $\mathcal{F}$ and/or $\mathcal{C}$. Methods in this direction could focus on physical world attacks that exploit camera properties, like the rolling shutter effect [134, 135]; rely on larger image disruptions like shadows [136]; or employ tools like the FR-disabling lasers used in Hong Kong in 2020 [72].

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 [137], making it likely that one or more unmodified photos of a user P will likely end up online, with or without P's knowledge. Given the widespread use of web scraping to collect FR reference images [5, 105], it is likely that one of these photos is already in a reference database.

Potential Directions. This stark reality has two implications for 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 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 [138, 139, 140]. 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: Privacy/utility/usability tradeoffs of AFR systems. The above paragraph raises an additional technical challenge of AFR design: balancing privacy, utility, and usability. There is a spectrum of ways to balance these. On one end are 3rd-party-adminstered AFR tools (c.f. stage ①), which have the high usability and utility but intrude on privacy to allow 3rd party data processing. On the other end are high-overhead tools like fully homomorphic encryption, which have provable privacy but limited utility/usability.

Where AFR tools should or can exist along this spectrum remains an open question for two reasons. First, we lack a deep understanding of how AFR users would prioritize these tradeoffs in practice. Prior studies show that users prefer privacy tools with minimal overhead [79, 80], but only one study has explored how or if these preferences change in the AFR setting [27]. Second, while many AFR tools have been proposed in recent years, the space of possible AFR designs remains sparsely populated. Consequently, it is worth considering whether this tradeoff is indeed fundamental, or if future AFR designs may evolve to accommodate all three.

TC 4: Face images don't change. A related, but distinct, challenge to TC 2 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. As our faces age, they remain recognizable as uniquely "us" to most humans and FR systems [141]. 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 $\mathbb{F}$ obtains P's protected face photo, they can try as many times as they want to break the protection [132]. If $\mathbb{F}$ ever succeeds, either in 1 month or 1 year, they "win" and P loses, because modern FR systems only need one clean picture in the reference database to identify a person [42]. 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 5: Lack of transparency of FR systems. The lack of transparency in how proprietary FR systems work 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. [44, 47]. While this may be sufficient to develop AFR tools that work well in the lab, researchers cannot perform comprehensive efficacy tests against proprietary systems.

Furthermore, AFR tool developers have no knowledge of how or if FR systems are adapting to evade AFR systems. The 2020 global FR market was valued at 3.86 billion US dollars [142], so FR stakeholders have ample resources to evolve as new AFR systems emerge. Even passive improvements to FR systems, such as new training methods or architectures, can overcome AFR protection and compromise user privacy [132]. Altogether, this lack of transparency means that AFR tools face an upward battle in the fight against unwanted FR.

11.2. Broader Social and Ethical Considerations

In addition to 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.

SEC 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 [40], to flourish. As a result, it is extremely difficult (if not impossible) for individuals to know when and where FR systems are deployed, as well as their capabilities.

This laissez-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, 3, 62, 143]. Government-sponsored FR may be *unwanted* but is not (necessarily) *unauthorized* under the status quo, and the legality of using AFR tools in this setting is ambiguous. To augment the confusion, systems like Clearview are used by law enforcement [5], further blurring the concept of *unauthorized* vs *unwanted* FR. As FR and AFR use increases, a clash over this issue seems almost inevitable.

SEC 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 [144, 145]. 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 [146]. Legitimate claims can be made by both sides. AFR researchers must be mindful of this tension and the potential consequences of their work.

SEC 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 $\mathbb{F}$ 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 from partial FR matches [60]. Furthermore, facial recognition systems misidentify people of color at higher rates [36, 147]. Recent work has found that AFR tools exhibit these same biases [148, 149]. The social impact of AFR tools requires urgent study.

12. Discussion

Related Surveys. Two surveys [34, 35], alluded to in §1, address topics similar to our work. Here, we provide an in-depth comparison to these and emphasize our unique contributions. [34] focuses on how a *service provider* can build a privacy-preserving FR service (see it's §II.D), while our work considers how *individual users* can use a privacy tool to defend themselves against *intrusive* FR services. Different from [34], our work provides a holistic view of what individual users can do to disrupt FR and a detailed discussion of challenges facing user-centric AFR solutions.

[35] also discusses methods a service provider could use to preserve privacy in a video surveillance setting.Furthermore, since [35] focuses on video surveillance, rather than image-based FR systems, it explores privacy preservation techniques for video-specific traits like gait, height, and clothing, which dilutes FR-specific content. [35] does address face de-identification in its §3.4.3, but focuses solutions for providers who wish to deanonymize *all* faces in their system (e.g. by averaging or blurring them), unlike our focus on protecting individuals from unwanted surveillance.

Concluding Thoughts. As facial recognition (FR) continues to grow in scale and ubiquity, we expect the demand for anti-facial recognition tools to continue to rise. 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.

Current AFR tools possess some, but not all, of the traits needed to successfully defeat unwanted FR in the real world. Many proposals leverage adversarial perturbations to evade FR models, either in the preprocessing ② or classification ⑤ stages. These are often effective in the short-term, but 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.

Acknowledgements. We thank our anonymous reviewers and shepherd for their insightful feedback. This work is supported in part by NSF grants CNS1949650, CNS-1923778, CNS-1705042, by C3.ai DTI, and by the DARPA GARD program. Emily Wenger is supported by a GFSD Fellowship and a Harvey Fellowship. Both Emily Wenger and Shawn Shan are also supported by Neubauer Fellowships at the University of Chicago. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

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," *ABC News*, April 30, 2021.

[5] K. Hill, "The Secretive Company that May End Privacy as We Know It," *The New York Times*, January 2020.

[6] ——, "Clearview AI's Facial Recognition App Called Illegal in Canada," *The New York Times*, 2021.

[7] US Customs and Border Control, "Collection of Biometric Data from Aliens Upon Entry to and Departure from the United States," *USCBP-2020-0062*, 2020.

[8] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT)," *NIST*, 2018.

[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*, November 2020.

[11] R. Chandran, "Fears for children's privacy as Delhi schools install facial recognition," *Reuters*, March 2021.

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

[13] A. Krishna, "IBM CEO's Letter to Congress on Racial Justice Reform," November 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," *CNBC*, November 2021.

[17] T. Li and M. S. Choi, "DeepBlur: A simple and effective method for natural image obfuscation," *arXiv preprint arXiv:2104.02655*, 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, "Ulixes: Facial recognition privacy with adversarial machine learning," *Proc. of PETs*, 2022.

[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*, 2021.

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

[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] B. Meden, P. Rot, P. Terhörst, N. Damer, A. Kuijper, W. J. Scheirer, A. Ross, P. Peer, and V. Štruc, “Privacy–enhancing face biometrics: A comprehensive survey,” *IEEE Tran. on Information Forensics and Security*, 2021.

[35] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, “Visual privacy protection methods: A survey,” *Expert Systems with Applications*, 2015.

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

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

[38] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, 1991.

[39] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE TPMI*, 1997.

[40] R. Heilweil, “The world’s scariest facial recognition company, explained,” *Vox.com*, May 2020.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, 2016.

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

[43] 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.

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

[45] 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.

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

[47] United States Government Accountability Office, “Facial Recognition Technology: Privacy and Accuracy Issues Related to Commercial Uses,” *GAO-20-522*, 2020.

[48] ——, “Facial Recognition Technology: Federal Law Enforcement Agencies Should Better Assess Privacy and Other Risks,” *GAO-21-518*, 2021.

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

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

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

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

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

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

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

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

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

[58] 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*, June 2018.

[59] 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.

[60] C. Garvie, “Garbage In, Garbage Out: Face Recognition on Flawed Data,” *Georgetown Law Center on Privacy and Technology*.

[61] IPVM, “Huawei/Megvii Uyghur Alarms,” December 2020.

[62] M. Mason, “Biometric Breakthrough: How CBP is Meeting Its Mandate and Keeping America Safe,” *U.S. Customs and Border Protection Website*.

[63] T. May and A. C. Chien, “Game Over: Chinese Company Deploys Facial Recognition to Limit Youths’ Play,” *New York Times*, July 2021.

[64] “Facial recognition tech fights coronavirus in Chinese city,” *France24 News*, July 2020.

[65] Visual Capitalist, “Mapped: The State of Facial Recognition Around the World,” 2020.

[66] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition,” *arXiv preprint arXiv:1607.08221*, 2016.

[67] 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.

[68] H.-W. Ng and S. Winkler, “A data-driven approach to cleaning large face datasets,” in *Proc. of ICIP*, 2014.

[69] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Learning Face Representation from Scratch,” *arXiv preprint arXiv:1411.7923*, 2014.

[70] United States Government Accountability Office, “Facial Recognition: CBP and TSA are Taking Steps to Implement Programs, but CBP Should Address Privacy and System Performance Issues,” *GAO-20-568*, 2020.

[71] “China State TV Exposes Wide Illegal Use Of Facial Recognition Cameras In Commercial Properties,” *China Money Network*, 2021.

[72] S. Mahtani and J. Hassan, “Hong Kong protesters are using lasers to distract and confuse. Police are shining lights right back.” *The Washington Post*, August 2019.

[73] B. Friedman and A. Ferguson, “Here’s a way forward on facial recognition,” *The New York Times*, October 2019.

[74] M. Devich-Cyril, “Defund facial recognition,” *The Atlantic*,

July 2020.

[75] A. Smith, “More Than Half of U.S. Adults Trust Law Enforcement to Use Facial Recognition Responsibly,” *Pew Research Center*, 2019.

[76] The Ada Lovelace Institute, “Beyond Face Value: Public Attitudes to Facial Recognition Technology,” 2019.

[77] 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.

[78] 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.

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

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

[81] 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.

[82] 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.

[83] J.-C. Chen, B.-H. Kung, K.-L. Hua, D. S. Tan *et al.*, “Naturalistic physical adversarial patch for object detectors,” in *Proc. of ICCV*, 2021.

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

[85] S. Yang, W. Wang, Y. Cheng, and J. Dong, “A Systematical Solution for Face De-identification,” in *Chinese Conference on Biometric Recognition*, 2021.

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

[87] S. Fu, F. He, Y. Liu, L. Shen, and D. Tao, “Robust unlearnable examples: Protecting data privacy against adversarial learning,” in *Proc. of ICLR*, 2021.

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

[89] 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.

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

[91] 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.

[92] 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.

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

[94] D. Deb, J. Zhang, and A. K. Jain, “Advfaces: Adversarial face synthesis,” in *Proc. of IJCB*, 2019.

[95] 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.

[96] I. Singh, S. Momiyama, K. Kakizaki, and T. Araki, “On Brightness Agnostic Adversarial Examples Against Face Recognition Systems,” in *BIOSIG*. IEEE, 2021.

[97] X. Yang, Y. Dong, T. Pang, H. Su, J. Zhu, Y. Chen, and H. Xue, “Towards Face Encryption by Generating Adversarial Identity Masks,” in *Proc. of ICCV*, 2021.

[98] 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.

[99] 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.

[100] K. Parikh, D. Singh, D. Yadav, and M. Rathod, “Detection of web scraping using machine learning,” *Open access international journal of Science and Engineering*, 2018.

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

[102] A. Haque and S. Singh, “Anti-scraping application development,” in *Proc. of ICACCI*, 2015.

[103] M. Clark, “Scraping by the Numbers,” *Facebook blog*, May 2021.

[104] 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.

[105] “Pimeyes.” [Online]. Available: https://pimeyes.com/en

[106] M. Rogoway, “Major Tech Company Using Facial Recognition to ID Workers,” *The Oregonian*, March 2020.

[107] 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*, May 2021.

[108] J. J. Roberts, “Walmart’s Use of Sci-fi Tech To Spot Shoplifters Raises Privacy Questions,” *Fortune*, November 2015.

[109] “How to protect your phone and identity at protests,” *Amnesty International*, January 2021.

[110] V. Ni and Y. Wang, “How to ’disappear’ on Happiness Avenue in Beijing,” *BBC*, November 2020.

[111] 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.

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

[113] 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.

[114] J. Bao, “Sparse Adversarial Attack to Object Detection,” *arXiv preprint arXiv:2012.13692*, 2020.

[115] 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.

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

[117] B. Liang, J. Li, and J. Huang, “We Can Always Catch You: Detecting Adversarial Patched Objects WITH or WITHOUT Signature,” *arXiv preprint arXiv:2106.05261*, 2021.

[118] J. Liu, A. Levine, C. P. Lau, R. Chellappa, and S. Feizi, “Segment and Complete: Defending Object Detectors against Adversarial Patch Attacks with Robust Patch Detection,” *arXiv preprint arXiv:2112.04532*, 2021.

[119] C. Xiang and P. Mittal, “DetectorGuard: Provably Securing Object Detectors against Localized Patch Hiding Attacks,” in *Proc. of CCS*, 2021.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

[125] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proc. of CCS*, 2017.

[126] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *arXiv preprint arXiv:1805.06605*, 2018.

[127] G. S. Dhillon, K. Azizzadenesheli, Z. C. Lipton, J. Bernstein, J. Kossaifi, A. Khanna, and A. Anandkumar, "Stochastic activation pruning for robust adversarial defense," *arXiv preprint arXiv:1803.01442*, 2018.

[128] N. Carlini and D. Wagner, "Magnet and 'efficient defenses against adversarial attacks' are not robust to adversarial examples," *arXiv preprint arXiv:1711.08478*, 2017.

[129] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. of ICML*, 2018.

[130] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *Proc. of NeurIPS*, 2020.

[131] 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.

[132] E. Radiya-Dixit and F. Tramer, "Data poisoning won't save you from facial recognition," *arXiv*, 2021.

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

[134] C.-K. Liang, L.-W. Chang, and H. H. Chen, "Analysis and compensation of rolling shutter effect," *IEEE Transactions on Image Processing*, 2008.

[135] A. Sayles, A. Hooda, M. Gupta, R. Chatterjee, and E. Fernandes, "Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect," in *Proc. of CVPR*, 2021.

[136] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," in *Proc. of CVPR*, 2022.

[137] P. Suciu, "A Photo Used To Be Worth A Thousand Words, But Thanks To Social Media Photos Have Lost Their Value," *Forbes online*, October 2019.

[138] 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.

[139] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. of IEEE S&P*, 2017.

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

[141] 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.

[142] Grandview Research, "Facial Recognition Market Size, Share and Trends Report, 2021 - 2028," 2021.

[143] A. Ziv, "This Israeli Face-recognition Startup is Secretly Tracking Palestinians," *Haaretz.com*, 2019.

[144] "Chinese Police Use Face Recognition to Find Child Abducted 30 Years Ago," *FindBiometrics.com*, May 2020.

[145] "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

[146] "International statement: End-to-end encryption and public safety," *United States Department of Justice, press release*, 2020.

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

[148] H. Rosenberg, B. Tang, K. Fawaz, and S. Jha, "Fairness Properties of Face Recognition and Obfuscation Systems," *arXiv preperint arXiv:2108.02707*, 2021.

[149] S. Qin, "Bias and Fairness of Evasion Attacks in Image Perturbation," *All Master's Theses, Digital Commons*, 2021.

[150] L. Mackenzie, "Surveillance state: how Gulf governments keep watch on us," *Wired Magazine*, 2021.

[151] L. Kayali, "How Facial Recognition is Taking Over a French City," *Politico*, 2019.

[152] C. Burt, "Kenyan police launch facial recognition on urban CCTV network," *Biometric Update*, September 2018.

[153] K. Pivcevic, "Police facial recognition use in Belarus, Greece, Myanmar raises rights, data privacy concerns," *Biometric Update*, March 2021.

[154] P. Fussey and D. Murray, "Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology," 2019.

[155] S. Jie, "China exports facial ID technology to Zimbabwe," *Global Times*, vol. 12, 2018.

[156] R. Mellen, "Buenos Aires is using facial recognition system that tracks child suspects, rights group says," *The Washington Post*, October 2020.

[157] H. Devlin, "We are hurtling towards a surveillance state: the rise of facial recognition technology," *The Guardian*, October 2019.

[158] P. Mozur, "Inside China's Dystopian Dreams: A.I., Shame and Lots of Cameras," *The New York Times*, 2018.

[159] C. Tan, "Malaysian police adopt Chinese AI surveillance technology," *Nikkei Asia*, 2018.

[160] A. Mascillno, "Facial recognition in schools: systems deployed in Europe and the US amid privacy concerns," *Biometric Update*, October 2021.

[161] M. Andrejevic and N. Selwyn, "Facial recognition technology in schools: Critical questions and concerns," *Learning, Media and Technology*, 2020.

[162] M. Z. Khan, "System soon to identify risky goods, individuals at borders," *Dawn News*, October 2020.

[163] A. Roussi, "Resisting the rise of facial recognition," *Nature*, 2021.

[164] H. Towney, "The retail stores you probably shop at that use facial-recognition technology," *Business Insider*, July 2021.

[165] U. Saiidi, "We went inside Alibaba's global headquarters.

Here's what we saw," *CNBC*, September 2019.
[166] "In-Car Biometric Technology For Human Interaction," *Hyundai Motor Group, blog article*, 2020.
[167] P. Lyon, "Subaru Forester Is First Mainstream Model To Offer Facial Recognition Technology," *Forbes*, April 2018.
[168] A. Smith, "JetBlue will test facial recognition for boarding," *CNN Business*, May 2017.
[169] "Could facial recognition be the future of airport security? Delta Air Lines is testing it out," *CBS News*, October 2021.
[170] D. Byler, "Because There Were Cameras, I Didn't Ask Any Questions," *China File*, December 2020.

Appendix

Tables 3 and 4 provide more context for the FR deployment and data collection content of §2.3. Table 3 lists known uses cases of FR systems by governments around the world. Table 4 lists known sources of labelled data for well-known large-scale FR systems, both public and private.

Location	Use Cases Reported	Countries/Companies
Public spaces	On-street surveillance	Bahrain [150], China [2], England [15], France [151], Kenya [152], Myanmar [153], Russia [4], UAE [150], UK [154], US [9], Zimbabwe [155]
	Criminal suspect identification	Argentina [156], Belarus [153], Brazil [157], China [158], Greece [153], Malaysia [159], US [1]
	School monitoring	Brazil [160], China [161], India [11], Russia [160], US [160]
	Border security	Israel [143], Pakistan [162], US [62]
	COVID lockdown enforcement	China [64], India [163], South Korea [163], Russia [4]
Privatized spaces	Catching shoplifters	Apple, Macy's, Lowe's [107, 164]
	Securing facility access	Alibaba [165], Intel [106]
	Tracking driver behavior	Hyundai [166], Subaru [167]
	Airline passenger check-in	JetBlue [168], Delta [169]

TABLE 3. EXAMPLE USE CASES OF FACIAL RECOGNITION.

Operator of FR system	Source of reference images
Clearview.ai	Social media photos [5]
PimEyes	(Public) online photos [105]
FBI F.A.C.E.S.	State drivers' license photos [1]
US Customs and Border Patrol	Passport photos [62]
Skynet (China)	National ID photos [3, 170]

TABLE 4. REPORTED REFERENCE IMAGE SOURCES