Personalized Privacy Protection Mask Against Unauthorized Facial Recognition

Ka-Ho Chow[1], Sihao Hu[2], Tiansheng Huang[2], and Ling Liu[2]

[1] The University of Hong Kong
[2] Georgia Institute of Technology
kachow@cs.hku.hk, {sihaohu,thuang374}@gatech.edu, lingliu@cc.gatech.edu

Abstract. Face recognition (FR) can be abused for privacy intrusion. Governments, private companies, or even individual attackers can collect facial images by web scraping to build an FR system identifying human faces without their consent. This paper introduces Chameleon, which learns to generate a user-centric personalized privacy protection mask, coined as P3-Mask, to protect facial images against unauthorized FR with three salient features. First, we use a cross-image optimization to generate one P3-Mask for each user instead of tailoring facial perturbation for each facial image of a user. It enables efficient and instant protection even for users with limited computing resources. Second, we incorporate a perceptibility optimization to preserve the visual quality of the protected facial images. Third, we strengthen the robustness of P3-Mask against unknown FR models by integrating focal diversity-optimized ensemble learning into the mask generation process. Extensive experiments on two benchmark datasets show that Chameleon outperforms three state-of-the-art methods with instant protection and minimal degradation of image quality. Furthermore, Chameleon enables cost-effective FR authorization using the P3-Mask as a personalized de-obfuscation key, and it demonstrates high resilience against adaptive adversaries.

1 Introduction

Face recognition (FR) has long been investigated due to its potential for enhancing security and convenience in various domains [22,28,29]. Many pretrained FR models are available online [6,7]. Once a face database (a.k.a. the gallery) with facial images for each person of interest is provided, these pretrained models can be used to recognize them [34].

While empowering many life-enriching applications, FR can be abused to cause serious privacy issues [2]. Privacy intruders can build a face database of victims of interest from publicly available facial images on the Internet by web scraping (Figure 1 (top)). By utilizing this database, the adversary can perform unauthorized FR of individual users for stalking victims [3], intruding on victims' privacy by flooding targeted ads [1], or facilitating criminals to commit identity fraud [5]. This is a real threat. For example, companies like Clearview [4] and PimEyes [8] have collected billions of online images and can recognize millions of citizens without their consent.

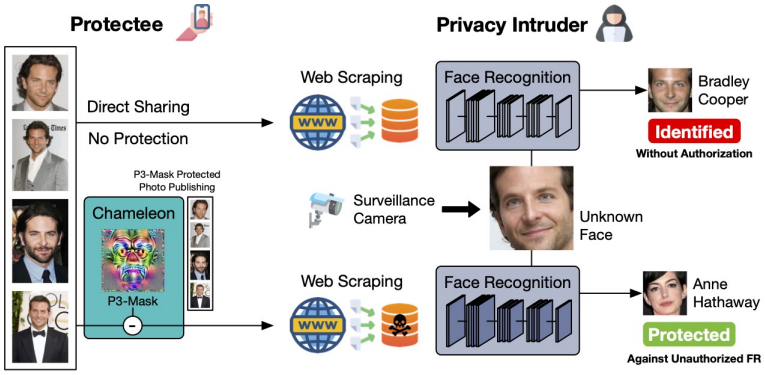


Fig. 1: (Top) Without protection, the privacy intruder can build a face database by web scraping and identify the unknown face. (Bottom) Chameleon learns the facial signature of the user (protectee) to generate a P3-Mask, which can be applied to protect any facial images before sharing them online against unauthorized FR.

Such a facial privacy threat has motivated the development of anti-FR technologies [36], which allow users (protectees) to preprocess their facial images before sharing them online. We argue that a competitive anti-FR solution should possess the following four properties. First, from the robustness perspective, the protected facial images should not be matched by an FR model to a query (probe) image with a correct identity (e.g., Bradley Cooper is recognized as Anne Hathaway in Figure 1 (bottom)), including those FR models possibly unknown during the preprocessing but are used by the privacy intruder. Second, from the efficiency perspective, the protection should be done instantly on any device with or without AI accelerators to maintain user experience and support equitable access to privacy protection. Third, from the perception usability perspective, the protected image should (i) preserve the visual quality rather than using excessive noise and (ii) be visually recognizable by humans to be the same person as the one in the original unprotected photo. Finally, the service usability for authorized FR providers is another important property. Existing approaches treat all FR models as unauthorized. We argue that anti-FR solutions should also allow users to grant FR permission to authorized models *cost-effectively*. This user-initiated de-obfuscation capability is critical to those, e.g., who publish their photos on a social media platform with privacy protection and, at the same time, wish to authorize the platform to perform some FR services (e.g., face tagging [32]) without sending and storing multiple versions of the same photo.

Based on the above objectives, we introduce Chameleon, a user-centric facial privacy protection system with three contributions. First, we develop an optimization algorithm to construct a Personalized Privacy Protection mask, coined as P3-Mask, for each user. The P3-Mask of a user can be used to protect any facial images of the same user with minimal impact on image quality, including those facial images unseen during the P3-Mask generation process. Second, it boosts the robustness of P3-Mask against unknown FR models by a princi-

pled approach, leveraging ensemble learning with models of high focal diversity. The per-user P3-Mask provides higher robustness against unknown FR models while preserving perception usability and the service usability for authorized FR providers. Chameleon users can utilize their own P3-Mask to obfuscate their facial images prior to public release, and the same P3-Mask can be used as a personalized de-obfuscation key to grant authorized FR service providers the ability to restore the facial signature of photos for correct recognition. Third, we conduct extensive experiments on two FR benchmark datasets [20,26] to analyze the shortcomings of state-of-the-art anti-FR solutions [31, 39, 40] and show that Chameleon can better protect against unknown FR models with a high success rate. The protection is lightweight and in real-time. It remains effective under adaptive privacy intruders deploying various strategies to counter Chameleon.

2 Related Work

Existing anti-FR approaches can be broadly classified into two categories. Synthesis-based approaches [11,16,19,21,33,41] use generative adversarial networks (GANs) [25] to synthesize a face to replace the original one for protection. While the synthetic faces look realistic, they appear to be strangers, not recognizable even by the users themselves, failing to meet the perception usability requirement.

Chameleon falls into the second category, which applies small changes to facial images [12]. Fawkes [31] formulates an untargeted attack to push the facial image away from the original location in the embedding space; TIP-IM [39] improves the image quality with MMD [9]; LowKey [10] improves the robustness under the image processing pipeline in an end-to-end ML system by incorporating it into the optimization process. These techniques can preserve the visual identity of protected faces. However, they both require iterative optimization for each image. Even for a GPU server, it can take over 100 seconds to perturb one facial image (Section 6.3). In contrast, Chameleon provides instant protection on any device, with protection even stronger than those spending minutes to find the best perturbation for each image. OPOM [40] attempts to improve efficiency by finding the embedding subspace enclosing facial images of the same person and generates a privacy protection mask to push any images away from it. However, it fails to maintain protection effectiveness when image processing operations are applied to protected images (Section 6.1), and the mask degrades the image quality much more significantly than Chameleon (Section 6.3).

3 Overview

To build an FR system with a pretrained model, the owner first collects a gallery dataset (face database) $\boldsymbol{\mathcal{D}}$, which includes the facial images of individuals of interest. Then, the pretrained FR model F is used to map each gallery image $\tilde{\boldsymbol{x}} \in \boldsymbol{\mathcal{D}}$ to the embedding $F(\tilde{\boldsymbol{x}})$, and those embeddings should form clusters corresponding to different people. When a facial image $\boldsymbol{x}$ of an unknown identity, called the probe image, is given, the FR system can use the same FR model F to

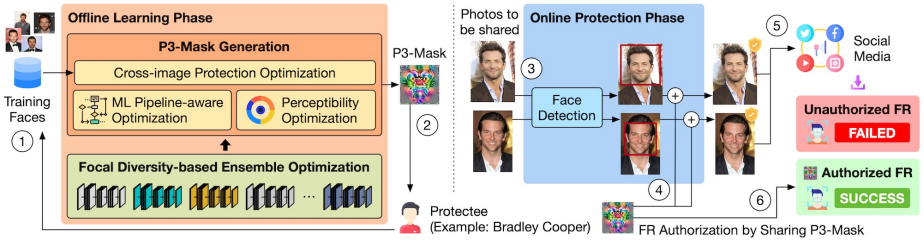


Fig. 2: An overview of Chameleon's two-phase workflow for offline learning to generate P3-Mask and online protection with P3-Mask for personalized facial signature masking. The P3-Mask can protect *any* facial images of the same person without further learning.

map it to an embedding $F(\boldsymbol{x})$ and use the identity of the nearest gallery image to be the identity of the unknown person. The identity, $\mathcal{FR}(\boldsymbol{x}; F, \boldsymbol{\mathcal{D}})$, of the probe image $\boldsymbol{x}$ given the FR model F and the gallery dataset $\boldsymbol{\mathcal{D}}$ is defined as:

$$\mathcal{FR}(\boldsymbol{x}; F, \boldsymbol{\mathcal{D}}) = \mathcal{I}\big(\arg\min_{\tilde{\boldsymbol{x}} \in \boldsymbol{\mathcal{D}}} \textsc{Dist}(F(\boldsymbol{x}), F(\tilde{\boldsymbol{x}}))\big), \tag{1}$$

where $\mathcal{I}(\tilde{\boldsymbol{x}})$ denotes the identity of the gallery image $\tilde{\boldsymbol{x}}$, which is known to the FR system, and $\textsc{Dist}(\cdot, \cdot)$ is a distance function such as Euclidean distance.

Threat Model. We consider a threat model where a privacy intruder conducts web scraping to collect images containing citizens' faces. Web scraping is large-scale and untargeted. Many citizens are included in the face database, and the privacy intruder may not know them. Given a probe image taken by, e.g., a stalker's camera, the privacy intruder uses an FR model on the face database to search for matched image(s). Then, the privacy intruder can analyze the associated metadata, such as the web page from where it was downloaded or even the exact identity. The goal of Chameleon is to allow the user to preprocess her facial images before posting them online. Even if scraped and included in the face database, they will not be matched to a probe image of her.

3.1 Chameleon Design

Chameleon performs preprocessing of the user's photos before sharing them online by applying the user-specific P3-Mask. This mask is generated offline for a Chameleon user. Figure 2 provides an overview of Chameleon's workflow.

Offline Learning Phase. Chameleon learns the unique facial signature of a user from a few facial images of her using the P3-Mask generator (Section 4), as shown in Figure 2 ①. P3-Mask is designed to apply directly to any facial images of the same user, including those unseen during offline learning, and is robust against lossy image processing operations in an end-to-end ML system without compromising the image quality. We use a focal diversity-optimized ensemble learning method to find a team of FR models such that the P3-Mask generated against them has strong robustness and is effective in countering other FR models that are unknown during offline learning.

Online Protection Phase. After the offline learning, ② the P3-Mask can be sent to the user's local device (e.g., the mobile phone). For a photo of the user to be protected, ③ a lightweight face detection model, such as MediaPipe [23], can be used by the Chameleon user running on the user's device to locate the face region and ④ instantly protect it with the P3-Mask before the user shares it online ⑤. The user can also authorize some trusted entities to conduct FR on their shared photos protected with P3-Mask. The user-specific de-obfuscation key ⑥ allows the protected photos to be restored for authorized FR.

4 P3-Mask Generation

For a user $\mathcal{P}$, Chameleon generates a P3-Mask by learning the facial signature $\boldsymbol{\mathcal{M}}_{\mathcal{P}}$ from a set of facial images the user provides. Several key factors need to be accomplished. (1) We need to promote cross-image protection as an optimization objective to generate the most representative facial signature applicable to protect any facial images of the user $\mathcal{P}$, including those unknown ones during the generation process (Section 4.1). (2) We need to control the amount of perturbation introduced by the P3-Mask. When applied to an unprotected image by removing the learned facial signature, it ensures minimal perception loss to preserve the visual quality (Section 4.2). (3) We need to keep the protected image far away from the original image in the embedding space. Multiple FR models should be used to generate alternative embedding representations to enhance generalizability against unknown models (Section 4.3).

4.1 Cross-image Protection Optimization

The cross-image protection capability comes from iterative learning on a set of training images the user provides. The idea is to keep fine-tuning the P3-Mask so that it can offer protection simultaneously to training images while preserving image quality. At the t-th iteration, it samples a mini-batch $\boldsymbol{\mathcal{B}}$ from the training dataset $\boldsymbol{\Omega}$ containing photos of user $\mathcal{P}$. Then, we find the modification to the P3-Mask that can lead to better protection on each image in the mini-batch with better visual quality with the following operations:

$$\boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t+1} = \textsc{Clip}_{[-\epsilon,\epsilon]}(\boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t} - \eta \textsc{Sign}(\frac{1}{|\boldsymbol{\mathcal{B}}|}\sum_{\mathcal{X}\in\boldsymbol{\mathcal{B}}} \nabla_{\boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}} \mathcal{L}(\mathcal{X}, \boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}; \boldsymbol{\mathcal{T}}, \omega))). \tag{2}$$

Specifically, for each image $\mathcal{X} \in \boldsymbol{\mathcal{B}}$, we use the P3-Mask learned up to the current iteration, $\boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}$, to protect it and compute the loss designed with dual goals:

$$\mathcal{L}(\mathcal{X}, \boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}; \boldsymbol{\mathcal{T}}, \omega) = \mathcal{L}_{\text{Protect}}(\mathcal{X}, \boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}; \boldsymbol{\mathcal{T}}) + \mathcal{L}_{\text{Percept}}(\mathcal{X}, \boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}; \omega). \tag{3}$$

It uses $\mathcal{L}_{\text{Protect}}$ to learn how to adjust the P3-Mask $\boldsymbol{\mathcal{M}}_{\mathcal{P}}^{t}$ from the previous iteration to better protect against a team of pre-selected FR models $\boldsymbol{\mathcal{T}}$ and $\mathcal{L}_{\text{Percept}}$ to preserve the image quality controlled by ω. We use the $\textsc{Sign}$ of the gradients to update the P3-Mask with a learning rate η. A clipping function $\textsc{Clip}_{[-\epsilon,\epsilon]}$ is applied to ensure the changes made by the P3-Mask on the facial

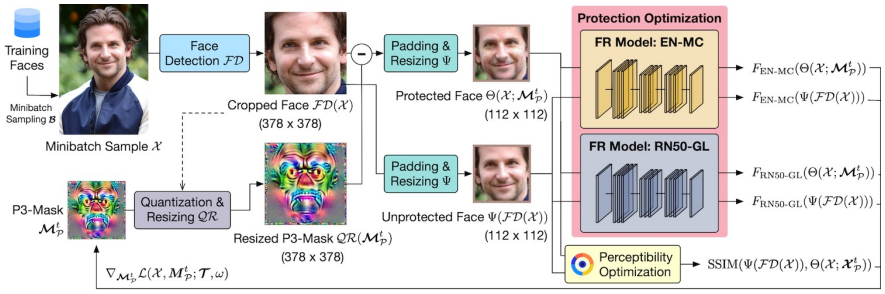


Fig. 3: The iterative generation of P3-Mask for a user. Chameleon goes through multiple facial images of the same user to optimize a P3-Mask with ML pipeline awareness.

image are bounded by ϵ. Such cross-image protection is not limited to those in the training dataset $\boldsymbol{\Omega}$ but also unseen images of the same user. We next discuss the design of $\mathcal{L}_{\text{Protect}}$, and $\mathcal{L}_{\text{Percept}}$ will be detailed in Section 4.2.

For privacy protection, P3-Mask maximizes the distance between the protected and unprotected images in the embedding spaces of a team $\boldsymbol{\mathcal{T}}$ of preselected FR models. We incorporate image processing operations into the optimization to avoid those lossy operations degrading P3-Mask's protection. Consider the raw training image $\mathcal{X}$ in the mini-batch in Figure 3. We first conduct face detection to produce the cropped face $\mathcal{FD}(\mathcal{X})$. Quantization and resizing are executed on the P3-Mask to match the resolution of the cropped face, denoted by $\mathcal{QR}(\boldsymbol{\mathcal{M}}^t_\mathcal{P})$. Then, we can apply the mask to the face for protection:

$$\Theta(\mathcal{X};\boldsymbol{\mathcal{M}}^t_\mathcal{P}) = \Psi(\textsc{Clip}_{[0,255]}(\mathcal{FD}(\mathcal{X}) - \mathcal{QR}(\boldsymbol{\mathcal{M}}^t_\mathcal{P}))), \tag{4}$$

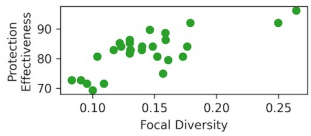
where Ψ is a padding and resizing function to meet the input resolution requirement of an FR model (e.g., (112×112) in ArcFace [17]). Note that to simplify the notation, we removed certain arguments without causing confusion.

The protection optimization in P3-Mask maximizes the average arccosine distance [17], ArcCos, between the embedding of the protected face $F(\Theta(\mathcal{X};\boldsymbol{\mathcal{M}}^t_\mathcal{P}))$ and the embedding of its unprotected counterpart $F(\Psi(\mathcal{FD}(\mathcal{X})))$ extracted by every FR model $F \in \boldsymbol{\mathcal{T}}$ in the team:

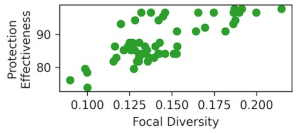
$$\mathcal{L}_{\text{Protect}}(\mathcal{X},\boldsymbol{\mathcal{M}}^t_\mathcal{P};\boldsymbol{\mathcal{T}}) = \frac{-1}{|\boldsymbol{\mathcal{T}}|}\sum_{F\in\boldsymbol{\mathcal{T}}} \textsc{ArcCos}(F(\Psi(\mathcal{FD}(\mathcal{X}))), F(\Theta(\mathcal{X};\boldsymbol{\mathcal{M}}^t_\mathcal{P}))). \tag{5}$$

4.2 Perceptibility Optimization

Preserving the visual quality of the facial image is necessary to maintain the usability of the protected image in practice. Hence, we incorporate a perceptibility optimization term in Equation 3. The idea is to minimize the perceptual difference between unprotected and protected images while learning a user's facial signature, such that removing the facial signature from a given facial image will have minimal impact on the visual quality of the protected version of the original image. We use the structural similarity (SSIM) [35] to capture perceptual



(a) Two-model: 28 Options



(b) Three-model: 56 Options

Fig. 4: Focal diversity provides a strong indicator of protection effectiveness for selecting an ensemble.

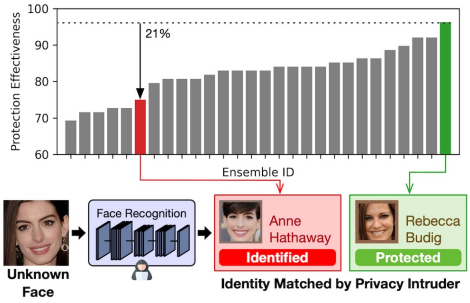


Fig. 5: Out of 28 options of two-model ensembles from a pool of eight models, the ensemble selected by our approach (green) leads to much better protection than a randomly selected one (red).

differences, as it has been shown to align with the human perception system:

$$\mathcal{L}_{\text{Percept}}(\mathcal{X}, \boldsymbol{\mathcal{M}}^t_{\mathcal{P}}; \omega) = \lambda_{\text{SSIM}} \max\left[\frac{1 - \text{SSIM}(\Psi(\mathcal{FD}(\mathcal{X})), \Theta(\mathcal{X}; \boldsymbol{\mathcal{M}}^t_{\mathcal{P}}))}{2} - \omega, 0\right]. \quad (6)$$

The parameters, ω and λ_{SSIM}, enable two features. ω controls the SSIM degradation. Any SSIM degradation greater than 2ω will cause the term to be non-zero, making the optimization adjust the P3-Mask reducing its impact on image quality. λ_{SSIM} balances the importance of privacy protection and perceptibility. We use dynamic scheduling [31] such that it will be adjusted automatically.

4.3 Focal Diversity Ensemble Optimization

For each original image and its protected version by removing the facial signature learned up to the current iteration, we use two FR models in Figure 3 to generate two embeddings, which serve as alternative channels to learn a high-quality facial signature. Choosing FR models that best complement each other can achieve better protection than randomly selected FR models. We use an ensemble optimization method to select the best k-model ensemble among a pool of N FR models. The key idea is to find an ensemble with members making decorrelated mistakes. To quantify this behavior, we leverage the focal diversity framework [13,14,37]. As shown in Figure 4, the higher the focal diversity of the ensemble (a green dot), the stronger the P3-Mask will be in protecting against other FR models *not* used during the P3-Mask generation process.

Given a collection of N FR models $\{F_1, ..., F_N\}$, we first identify the negative samples for each FR model F by locating validation images that F fails to recognize their true identity. To rank ensembles of size S, we enumerate all $\binom{N}{S}$ combinations. For each combination (ensemble) $\boldsymbol{\mathcal{T}}$, we consider each member to be the focal model F_{focal} and use its negative samples to statistically estimate the level of negative correlation $\lambda_{\text{focal}}(\boldsymbol{\mathcal{T}}; F_{\text{focal}})$ between F_{focal} and the remaining models in $\boldsymbol{\mathcal{T}}$, computed by measuring the degree of disagreements using the

generalized non-pairwise measure [27]. The same procedure is repeated by considering each member in the ensemble as the focal model, and the focal diversity of the ensemble $\boldsymbol{\mathcal{T}}$ is finalized as $d_{\text{focal}}(\boldsymbol{\mathcal{T}}) = \frac{1}{S}\sum_{F_{\text{focal}} \in \boldsymbol{\mathcal{T}}}[1 - \lambda_{\text{focal}}(\boldsymbol{\mathcal{T}}; F_{\text{focal}})]$. Given a team size S, the ensemble with the highest focal diversity can be deployed. As shown in Figure 5, the selected two-model team (green bar) is indeed the one leading to the strongest protection among all 28 options, which is over 20% stronger than a randomly selected ensemble (red bar). Note that we only need to analyze the failures of individual FR models, which is significantly more efficient than generating P3-Mask for each option and conducting evaluation.

5 Online Image Protection and Refinement

The user $\mathcal{P}$ obtains her P3-Mask $\boldsymbol{\mathcal{M}}_{\mathcal{P}}$ from Chameleon. Given an image $\mathcal{X}$, she can obfuscate her facial identity on her device by applying the P3-Mask:

$$\textsc{Mask}(\mathcal{X}; \boldsymbol{\mathcal{M}}_{\mathcal{P}}) = \textsc{Clip}_{[0,255]}(\mathcal{FD}(\mathcal{X}) - \mathcal{QR}(\boldsymbol{\mathcal{M}}_{\mathcal{P}})), \tag{7}$$

which is the protected facial region to be put in the original image $\mathcal{X}$ to produce the protected version $\mathcal{X}'$ for sharing.

Chameleon supports authorizing trusted third parties to conduct FR on a user's facial images in a space-time efficient manner without transmitting and storing multiple versions of the same photo (one protected for public sharing and one unprotected for internal FR). By sharing the P3-Mask $\boldsymbol{\mathcal{M}}_{\mathcal{P}}$ as the key, the third parties can de-obfuscate the facial signature by unmasking:

$$\textsc{Unmask}(\mathcal{X}'; \boldsymbol{\mathcal{M}}_{\mathcal{P}}) = \textsc{Clip}_{[0,255]}(\mathcal{FD}(\mathcal{X}') + \mathcal{QR}(\boldsymbol{\mathcal{M}}_{\mathcal{P}})). \tag{8}$$

In Section 6.2, we will show two properties of this process: (i) the signature-restored photos can be used for FR with no accuracy degradation, and (ii) the protection can only be done by the P3-Mask of the same person in the facial image, and the restoration is successful only if the same key is used.

6 Experimental Evaluation

We conduct experiments on FaceScrub [26] and LFW [20]. To provide a detailed analysis, ten celebrities are randomly selected as users on FaceScrub (Table 1). For each user, we split their facial images into three parts: (i) 10% are used as probe images (2nd column), (ii) 70% are used as gallery images and are used by Chameleon to train the mask (3rd column), and (iii) 20% are also used as gallery images but unseen during the training process (4th column). The last part is crucial to evaluate the protection of unseen images of the user. In total, we have 123 facial images with "unknown identities." Following relevant works [31, 40], all facial images of other people are included in the gallery dataset as noise. The same splitting method is also used for LFW. By default, the results reported in this paper focus on FaceScrub due to similar observations on LFW. The source code is available at `https://github.com/git-disl/Chameleon`.

Table 1: Ten example celebrities (users) on FaceScrub for analysis.

Name	Probe Images	Gallery Images		
		Seen	Unseen	Total
Morena Baccarin	11	82	23	105
Bradley Cooper	12	89	25	114
America Ferrera	16	113	32	145
Gerard Butler	11	82	24	106
Eva Longoria	12	91	26	117
Melissa Egan	12	87	24	111
Kim Cattrall	13	96	27	123
Allison Janney	12	89	26	115
Roma Downey	11	78	22	100
Steve Carell	13	96	28	124
Others (Noise)	/	/	/	48368
Total Number of Gallery Images:				49528

Table 2: The collection of publicly available pretrained FR models.

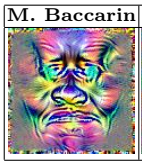
Model ID	Neural Arch.	Training Dataset	FR Acc.
EN-MC	EfficientNet	MS-Celeb-1M	94.94%
RN50-GL	ResNet50	Glint360K	96.68%
RN50-MC	ResNet50	MS-Celeb-1M	89.25%
RN50-VF	ResNet50	VGG-Face2	94.30%
RN50-WF	ResNet50	WebFace600K	96.72%
RN18-MC	ResNet18	MS-Celeb-1M	82.64%
RN34-MC	ResNet34	MS-Celeb-1M	84.49%
RN100-MC	ResNet100	MS-Celeb-1M	91.85%

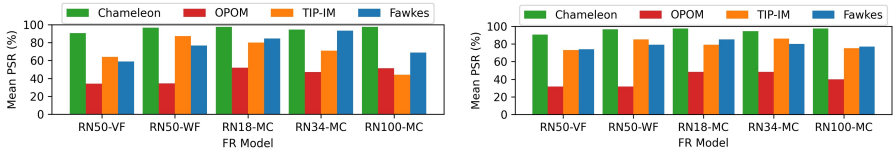
Public clouds (e.g., Azure) now require manual approval to use their FR services. Hence, we believe that privacy intruders will opt for deploying pretrained FR models, as many high-quality ones are available on the Internet and can be used out of the box. Due to the large number of possible FR algorithms and architectures, we first focus on the eight FR models listed in Table 2, which are based on ArcFace [17], the state-of-the-art FR algorithm, with varying neural architectures and training datasets. Chameleon can be easily extended to incorporate any FR models by simply adding them to the collection. Nonetheless, in Section 6.4, we will show that thanks to the focal diversity-optimized teaming, the P3-Mask generated from a collection of ArcFace-only models can also be effective in protecting against privacy intruders using models of other FR algorithms, such as FaceNet [30] and MagFace [24]. By default, we use the two-model team (EN-MC, RN50-GL) with the highest focal diversity (Section 4.3). The P3-Mask is trained for 50 epochs on NVIDIA RTX 2080 SUPER GPU with $\eta = 0.001$, $|\boldsymbol{\mathcal{B}}| = 4$, $\epsilon = 0.063$, and $\omega = 0.03$. We provide details for reproducibility and additional results in the appendix.

6.1 Protection Effectiveness

Table 3 shows the personalized mask for seven users on FaceScrub (other users are available in the appendix). Different users have distinct facial signatures (P3-Mask) learned by Chameleon. We apply these masks to their respective gallery

Table 3: The P3-Mask (rescaled) for seven users on FaceScrub. Chameleon learns the distinct facial signature for each of them.

M. Baccarin	B. Cooper	A. Ferrera	G. Butler	E. Longoria	M. Egan	K. Cattrall
						



(a) Dataset: FaceScrub **(b)** Dataset: LFW

Fig. 6: Protection effectiveness on two benchmarks using Chameleon, OPOM [40], TIP-IM [39], and Fawkes [31] against FR models unknown to them.

images and test the FR accuracy using probe images. All probe images can be correctly identified when no protection mechanism is used. Under Chameleon, the probe images should be matched to an incorrect identity, resulting in a low FR accuracy. Hence, we define an evaluation metric, Protection Success Rate (PSR), to be (100 − FR Accuacy) reported in percentages.

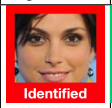
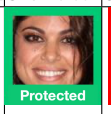
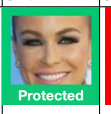
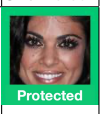
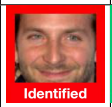
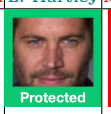
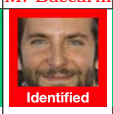
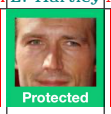
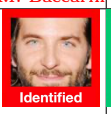
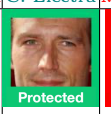
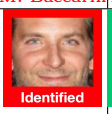
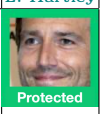
Chameleon can protect against unknown FR models. We compare the cross-model protection of Chameleon with OPOM [40], the only anti-FR approach generating one privacy protection mask for each person as Chameleon, and both TIP-IM [39] and Fawkes [31], which conduct per-image optimization. For a fair comparison, all methods are set with the same perturbation budget. Figure 6 summarizes the results on both datasets. All five FR models are *unknown* to both protection mechanisms. We make two observations. First, Chameleon consistently outperforms OPOM by a large margin. We found that OPOM masks do not transfer well, especially considering the end-to-end ML pipeline. Its PSR can drop by 33% because of those lossy operations. Second, Chameleon can be as competitive as those methods conducting per-image optimization. TIP-IM and Fawkes spend minutes to optimize the protection for *each* image. Still, Chameleon outperforms them and can be applied instantly.

Table 4 reports the detailed PSR for each user Chameleon achieves when the privacy intruder uses different FR models (2nd to 9th columns). We make two observations. First, when the privacy intruder uses an FR model known to the generation process, a PSR of 100% can be consistently achieved (2nd to 3rd columns). Second, Chameleon can protect against unknown FR models (4th to 9th columns). Even when the FR model used by the intruder is trained on a different dataset (i.e., RN50-VF and RN50-WF) or with a different neural architecture (i.e., RN18-MC, RN34-MC, and RN100-MC) than any FR models known by Chameleon, it still provides a PSR over 90.65%, meaning that facial images of a user are matched to gallery images of a different person. To demonstrate such a mismatch, in Table 5, we show the probe images from two users and their most similar gallery images found by four unknown FR models with two settings: (i) the “Unprotected" scenario where no one employs protection and (ii) the “Chameleon" scenario where the intruder scraped photos protected by our solution. Taking M. Baccarin as an example, the most similar gallery images found in the unprotected scenario belong to her. In contrast, when Chameleon is used, the same probe image is misidentified, e.g., as L. Hartley by RN50-MC.

Table 4: Using the two-model team with high focal diversity (2nd and 3rd columns), Chameleon can offer privacy protection even when the privacy intruder uses FR models unseen during the mask generation process (4th to 9th columns).

Name	Protection Success Rate - PSR (%)									
	EN-MC	RN50-GL	RN50-MC	RN50-VF	RN50-WF	RN18-MC	RN34-MC	RN100-MC	Mean	Std
M. Baccarin	100.00	100.00	100.00	72.73	90.91	100.00	90.91	100.00	94.32	9.64
B. Cooper	100.00	100.00	83.33	100.00	100.00	100.00	91.67	91.67	95.83	6.30
A. Ferrera	100.00	100.00	93.75	75.00	100.00	100.00	93.75	100.00	95.31	8.68
G. Butler	100.00	100.00	90.91	100.00	100.00	90.91	90.91	100.00	96.59	4.70
E. Longoria	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00
M. Egan	100.00	100.00	100.00	100.00	100.00	100.00	100.00	91.67	98.96	2.95
K. Cattrall	100.00	100.00	100.00	84.62	100.00	100.00	100.00	100.00	98.08	5.44
A. Janney	100.00	100.00	91.67	100.00	91.67	100.00	100.00	100.00	97.92	3.86
R. Downey	100.00	100.00	100.00	72.73	90.91	100.00	90.91	100.00	94.32	9.64
S. Carell	100.00	100.00	84.62	92.31	92.31	92.31	76.92	100.00	92.31	8.22
Mean	100.00	100.00	93.52	90.65	96.58	97.41	94.42	97.42		
Std	0.00	0.00	6.40	11.16	4.43	4.18	7.41	4.15		

Table 5: The most similar gallery images on FaceScrub found by different FR models unknown to Chameleon.

Probe Image	The Most Similar Gallery Image & Its Identity Found by Unseen FR Models							
	RN50-MC		RN50-VF		RN34-MC		RN100-MC	
	Unprotected	Chameleon	Unprotected	Chameleon	Unprotected	Chameleon	Unprotected	Chameleon
	 Identified	 Protected	 Identified	 Protected	 Identified	 Protected	 Identified	 Protected
M. Baccarin	M. Baccarin	L. Hartley	M. Baccarin	L. Hartley	M. Baccarin	C. Electra	M. Baccarin	L. Hartley
	 Identified	 Protected	 Identified	 Protected	 Identified	 Protected	 Identified	 Protected
B. Cooper	B. Cooper	P. Walker	B. Cooper	M. Vartan	B. Cooper	M. Vartan	B. Cooper	M. Vartan

6.2 FR Service Usability

Chameleon allows users to grant FR permission to trusted third parties by sharing their P3-Mask to de-obfuscate the protected image. There is no need to transmit and store an unprotected photo for internal use and a protected one for public visibility, doubling network and storage costs. Table 6 shows the results in FR accuracy. First, using the correct mask (i.e., the one used to protect the images) to reverse the protection can restore FR accuracy. Taking M. Baccarin as an example, the FR accuracy increases from 6.82% before unmasking (b) to 100% after unmasking (c). Second, using an incorrect mask for unmasking cannot restore the FR accuracy and can lead to even worse performance. It drops from 6.82% to 1.14% after unmasking (d).

6.3 Protection Cost Analysis

Speed and Resources. Figure 7 compares Chameleon with OPOM, TIP-IM, and Fawkes regarding the protection time per face, compute, and storage costs.

Table 6: The protection by Chameleon can reduce FR accuracy against FR models (b), but the user can provide the P3-Mask used to protect her images as the key for the authorized third party to reverse the protection process (c), which leads to restored FR performance. However, an incorrect key cannot restore FR and may worsen it (d).

Scenario	Face Recognition Accuracy (%)									
	EN-MC	RN50-GL	RN50-MC	RN50-VF	RN50-WF	RN18-MC	RN34-MC	RN100-MC	Mean	Std
User: M. Baccarin										
(a) No Protection	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00
(b) Protected	0.00	0.00	9.09	18.18	9.09	9.09	0.00	9.09	6.82	6.43
(c) Correctly Unmasked	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00
(d) Incorrectly Unmasked	0.00	0.00	0.00	0.00	9.09	0.00	0.00	0.00	1.14	3.21
User: B. Cooper										
(a) No Protection	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00
(b) Protected	0.00	0.00	16.67	0.00	0.00	0.00	8.33	8.33	4.17	6.30
(c) Correctly Unmasked	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0.00
(d) Incorrectly Unmasked	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

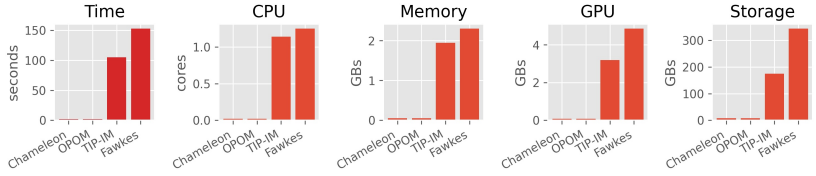


Fig. 7: Chameleon offers much better privacy protection than OPOM without sacrificing speed and resource usages. It can protect instantly and require negligible compute resources as OPOM, while TIP-IM and Fawkes are slow and costly.

Chameleon and OPOM produce masks that are applicable to any facial image of the same user. The protection can be completed in 0.0076 seconds. Note that Chameleon is significantly more effective than OPOM, as described in Figure 6. Instead, TIP-IM and Fawkes require per-image optimization and take 105.12 seconds to complete the protection with GPUs for acceleration and storage for FR models. Chameleon only needs to conduct simple arithmetic operations. Only the P3-Mask needs to be stored on the user device. Hence, it can be deployed with instant protection even on an edge device with weak computing power.

Image Quality. Chameleon can well preserve image quality. Table 7 provides two examples for four users, contrasting the facial images with no protection (3rd and 5th columns) with the ones with their facial signature removed by Chameleon (4th and 6th columns). They capture different variations of facial images: M. Baccarin's images have different lighting conditions, B. Cooper's have different face sizes, E. Longoria's have different postures, and G. Bulter's have different expressions. The protected images are visually similar to the unprotected counterparts, but they will not be matched to the clean images of the corresponding person. Indeed, Chameleon can generate higher-quality images than existing methods. As reported in Table 8, Chameleon not only offers better protection (Figure 6) with a similar cost (Figure 7) as OPOM but also achieves much better image quality measured in SSIM (0.9493 vs 0.8839).

Table 7: The P3-Mask can be applied to any facial images of the same person while preserving image quality.

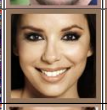
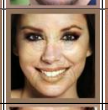
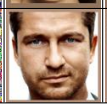
User	Mask	Sample Image 1		Sample Image 2	
		No Protection	Protected	No Protection	Protected
M. Baccarin					
B. Cooper					
E. Longoria					
G. Image					

Table 8: Chameleon offers much better protection while maintaining image quality.

	Example
Original **SSIM**: 1.0000	
Chameleon **SSIM**: 0.9493	
OPOM **SSIM**: 0.8839	
TIP-IM **SSIM**: 0.8850	
Fawkes **SSIM**: 0.9612	

6.4 Focal Diversity-based Teaming

Chameleon automatically selects a high-quality team for deployment with a specified budget. Table 9 compares the detailed PSR using the most diverse and the least diverse teams with (a) two and (b) three models. In both cases, the most diverse team outperforms the least diverse one, which is only effective when the privacy intruder uses the FR model known in the team. The selection of high-quality teams is non-trivial because we observe that composing a team of FR models with the highest FR accuracy is not always a good option, and a team of models having different neural architectures is not always the top priority [13].

In Figure 8, we further show that the P3-Mask generated from a carefully-chosen team can protect against FR models of unknown algorithms. Specifically,

Table 9: The most diverse and the least diverse teams with (a) two or (b) three models identified by our focal diversity-based teaming method. The most diverse teams offer significantly better protection in terms of protection success rates.

Protection Team	Protection Success Rate - PSR (%)									
	EN-MC	RN50-GL	RN50-MC	RN50-VF	RN50-WF	RN18-MC	RN34-MC	RN100-MC	Mean	Std
(a) Two-model Teams										
Most Diverse: (EN-MC, RN50-GL)	100.00	100.00	93.52	90.65	96.58	97.41	94.42	97.42	**96.25**	**3.23**
Least Diverse: (RN18-MC, RN34-MC)	11.38	60.16	100.00	55.28	73.17	100.00	100.00	80.49	72.56	28.54
(b) Three-model Teams										
Most Diverse: (EN-MC, RN50-GL, RN50-WF)	100.00	100.00	100.00	100.00	92.68	93.50	100.00	100.00	**98.27**	**3.00**
Least Diverse: (RN18-MC, RN34-MC, RN50-MC)	2.44	78.86	100.00	52.03	84.55	100.00	100.00	93.50	76.42	31.82

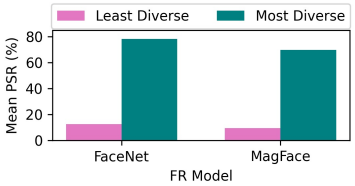


Fig. 8: The most diverse teams allow Chameleon to be effective against FR models of unknown algorithms.

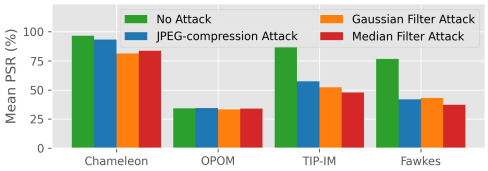


Fig. 9: The PSR of Chameleon against adaptive intruders using different strategies to wash out the patterns in protected images.

we consider the most and the least diverse three-model teams in Table 9(b). Even though both teams include only FR models based on ArcFace [17], the most diverse one can effectively protect against FR models based on FaceNet [30] or MagFace [24]. The protection effectiveness can be further strengthened by incorporating more FR models into the collection.

6.5 Chameleon Against Adaptive Adversaries

An adaptive adversary may run Chameleon to produce the P3-Mask for each person and de-obfuscate their photos. When probe images need to be identified, they will be matched against a “clean" face database. However, it is infeasible, as we show in Section 6.2, that the FR accuracy can only be restored when the masks used for obfuscation and de-obfuscation are identical. It is impossible to reproduce the same P3-Mask used by the user because it requires the same set of clean images and random states.

Alternatively, an adaptive adversary may wash out patterns introduced by Chameleon before performing FR. We consider three popular lightweight and task-agnostic methods studied in the adversarial example domain [18] in Figure 9, including (a) JPEG compression [15], (b) Gaussian Filter, and (d) Median Filter [38]. The results show that their effectiveness is limited on Chameleon. The consideration of the end-to-end ML pipeline increases the robustness of P3-Mask.

7 Conclusions

We have presented Chameleon against unauthorized FR. Chameleon generates a P3-Mask for each user with cross-image and perceptibility optimizations to offer (i) instant and lightweight protection on any facial images of the same user and (ii) preservation of image quality. Also, we have shown that P3-Mask enables cost-effective de-obfuscation for authorizing FR services. Such an authorization process can only be conducted by the one with the identical mask used for protection. Chameleon also features a focal diversity-optimized teaming method to select a high-quality FR team to generate P3-Mask with strong robustness against unknown FR models.

Acknowledgments

This research is partially sponsored by the NSF CISE grants 2302720, 2312758, 2038029, an IBM faculty award, and a grant from CISCO Edge AI program. It is part of the PhD dissertation of the first author, who graduated from Georgia Tech in Spring 2023. The first author acknowledges the support of the IBM PhD Fellowship in 2022-2023 and the support from the HKU-CS Start-up Fund.

References

1. Brazilian retailer quizzed over facial recognition tech. https://www.zdnet.com/article/brazilian-retailer-quizzed-over-facial-recognition-tech/ (2019)
2. Facial recognition in the united states: Privacy concerns and legal developments. https://www.asisonline.org/security-management-magazine/monthly-issues/security-technology/archive/2021/december/facial-recognition-in-the-us-privacy-concerns-and-legal-developments/ (2021)
3. The secretive company that might end privacy as we know it. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html (2021)
4. Clearview AI. https://www.clearview.ai/ (2023)
5. Facial recognition and identity risk. https://www.equifax.co.uk/resources/identity-protection/facial-recognition-and-identity-risk.html (2023)
6. HuggingFace. https://huggingface.co (2023)
7. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface (2023)
8. PimEyes. https://pimeyes.com/ (2023)
9. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics **22**(14), e49–e57 (2006)
10. Cherepanova, V., Goldblum, M., Foley, H., Duan, S., Dickerson, J., Taylor, G., Goldstein, T.: Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In: International Conference on Learning Representations (ICLR) (2021)
11. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8789–8797 (2018)
12. Chow, K.H., Hu, S., Huang, T., Ilhan, F., Wei, W., Liu, L.: Diversity-driven privacy protection masks against unauthorized face recognition. Proceedings on Privacy Enhancing Technologies **4**, 381–392 (2024)
13. Chow, K.H., Liu, L.: Robust object detection fusion against deception. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 2703–2713 (2021)
14. Chow, K.H., Liu, L.: Boosting object detection ensembles with error diversity. In: 2022 IEEE International Conference on Data Mining (ICDM). pp. 903–908. IEEE (2022)

15. Das, N., Shanbhogue, M., Chen, S.T., Hohman, F., Li, S., Chen, L., Kounavis, M.E., Chau, D.H.: Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 196–204 (2018)
16. Deb, D., Zhang, J., Jain, A.K.: Advfaces: Adversarial face synthesis. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2020)
17. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
18. Guo, C., Rana, M., Cisse, M., Van Der Maaten, L.: Countering adversarial images using input transformations. arXiv preprint arXiv:1711.00117 (2017)
19. Hu, S., Liu, X., Zhang, Y., Li, M., Zhang, L.Y., Jin, H., Wu, L.: Protecting facial privacy: generating adversarial identity masks via style-robust makeup transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15014–15023 (2022)
20. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
21. Li, T., Lin, L.: Anonymousnet: Natural face de-identification with measurable privacy. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
22. Liu, L., Zhou, B., Zou, Z., Yeh, S.C., Zheng, L.: A smart unstaffed retail shop based on artificial intelligence and iot. In: 2018 IEEE 23rd International workshop on computer aided modeling and design of communication links and networks (CAMAD). pp. 1–4. IEEE (2018)
23. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
24. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: A universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14225–14234 (2021)
25. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
26. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP). pp. 343–347. IEEE (2014)
27. Partridge, D., Krzanowski, W.: Software diversity: practical statistics for its measurement and exploitation. Information and software technology **39**(10), 707–717 (1997)
28. Pinto, N., Stone, Z., Zickler, T., Cox, D.: Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In: CVPR 2011 WORKSHOPS. pp. 35–42. IEEE (2011)
29. Sahani, M., Nanda, C., Sahu, A.K., Pattnaik, B.: Web-based online embedded door access control and home security system based on face recognition. In: 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]. pp. 1–6. IEEE (2015)
30. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)

31. Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., Zhao, B.Y.: Fawkes: Protecting privacy against unauthorized deep learning models. In: Proceedings of the 29th USENIX Security Symposium (2020)
32. Stone, Z., Zickler, T., Darrell, T.: Autotagging facebook: Social network context improves photo annotation. In: 2008 IEEE computer society conference on computer vision and pattern recognition workshops. pp. 1–8. IEEE (2008)
33. Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., Schiele, B.: A hybrid model for identity obfuscation by face replacement. In: Proceedings of the European conference on computer vision (ECCV). pp. 553–569 (2018)
34. Wang, M., Deng, W.: Deep face recognition: A survey. Neurocomputing **429**, 215–244 (2021)
35. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
36. Wenger, E., Shan, S., Zheng, H., Zhao, B.Y.: Sok: Anti-facial recognition technology. In: 2023 IEEE Symposium on Security and Privacy (SP). pp. 134–151 (2023)
37. Wu, Y., Liu, L., Xie, Z., Chow, K.H., Wei, W.: Boosting ensemble accuracy by revisiting ensemble diversity metrics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16469–16477 (2021)
38. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks (2018)
39. Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J., Chen, Y., Xue, H.: Towards face encryption by generating adversarial identity masks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3897–3907 (2021)
40. Zhong, Y., Deng, W.: Opom: Customized invisible cloak towards face privacy protection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
41. Zhu, Z.A., Lu, Y.Z., Chiang, C.K.: Generating adversarial examples by makeup attacks on face recognition. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 2516–2520. IEEE (2019)