Diversity-driven Privacy Protection Masks Against Unauthorized Face Recognition

Ka-Ho Chow
The University of Hong Kong
kachow@cs.hku.hk

Sihao Hu
Georgia Institute of Technology
sihaohu@gatech.edu

Tiansheng Huang
Georgia Institute of Technology
thuang374@gatech.edu

Fatih Ilhan
Georgia Institute of Technology
filhan@gatech.edu

Wenqi Wei
Georgia Institute of Technology
wenqiwei@gatech.edu

Ling Liu
Georgia Institute of Technology
lingliu@cc.gatech.edu

ABSTRACT

Face recognition (FR) technologies have enabled many life-enriching applications but have also opened doors for potential misuse. Governments, private companies, or even individuals can scrape the web, collect facial images, and build a face database to fuel the FR system to identify human faces without their consent. This paper introduces PMask to combat such a privacy threat against unauthorized FR. It provides a holistic approach to enable privacy-preserving sharing of facial images. PMask preprocesses the facial image and hides its unique facial signature through iterative optimization with dual goals: (i) minimizing the amount of noise to ensure high image quality and (ii) minimizing the perception loss between the privacy-protected face and the original face to ensure the face is recognizable to be the same person by humans. Extensive experiments are conducted on eight representative FR models to evaluate PMask against unauthorized FR. The results validate that PMask provides much stronger protection, introduces less perceptible changes to facial images, and runs faster than state-of-the-art methods to provide privacy protection with a better user experience.

1 INTRODUCTION

Face recognition (FR) technologies have made their way into our everyday lives [8, 26, 30, 31]. Many pretrained FR models are available online for free [6]. Once a face database (a.k.a. the gallery) with facial images for each person of interest is provided, the pretrained models can be used to recognize people effectively, even when they are unknown at the model training phase. While FR technologies have powered numerous life-enriching applications, misuse of FR can cause serious privacy intrusions [2]. For example, privacy intruders can conduct web scraping to build a face database out of publicly available images on the Internet. Then, by sending a facial image as a query against this gallery, privacy intruders can infer the identity of the person with high confidence, as shown in **Figure 1a**. For example, a private company, Clearview.ai [4], has already collected over 20 billion online images and can recognize millions of citizens without their consent. This is a real threat: stalkers can find out the footprints of their victims [3], retail stores may associate online browsing history with offline shopping behaviors for advertisement [1], and criminals may commit identity fraud [5].

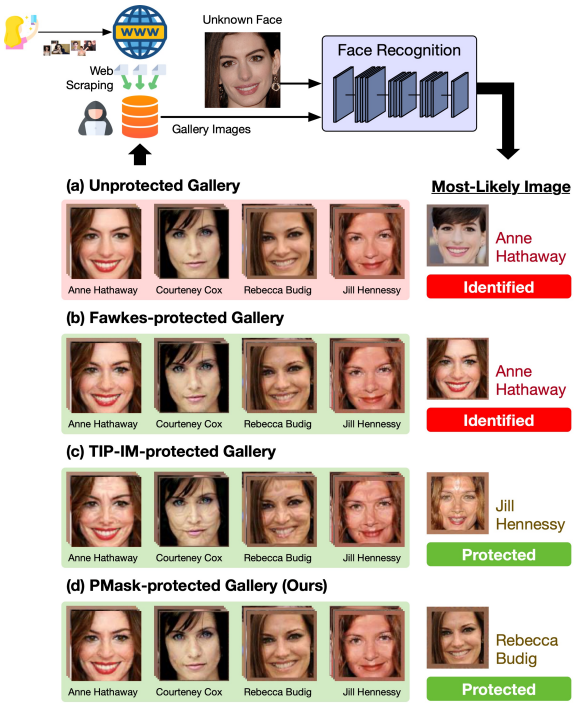


Figure 1: The privacy intruder scrapes the web to build a gallery for identifying unknown faces. Compared with Fawkes [33] and TIP-IM [42], PMask can provide better protection with a small noise injected into the gallery images.

To protect against unauthorized FR using scraped photos containing facial images of citizens, we propose PMask, a facial masking system for privacy-preserving facial image publishing. Before sharing a photo with facial images on the Internet, PMask will inject a privacy mask, which can effectively hide the true facial signature of the image while maintaining its visual perception quality. **Figure 1** illustrates the benefit of our proposed PMask with the unknown face as query input by comparing (d) the PMask-protected Gallery (ours) to (a) the Unprotected Gallery, (b) the Fawkes-protected Gallery [33], and (c) the TIP-IM-protected Gallery [42]. It is observed that for the two recent privacy protection systems, Fawkes and TIP-IM, one fails to protect privacy (b), while another adds too

Proceedings on Privacy Enhancing Technologies 2024(4), 381–392

https://doi.org/10.56553/popets-2024-0122

much noise to the gallery images (c). Even though it succeeds in protecting the input image from correct recognition, the preservation of the visual perception quality fails. In comparison, PMask can generate the privacy mask, which not only ensures that it can effectively hide the unique facial signature of gallery images but also guarantees that the noise added during privacy mask generation is minimal and thus preserves the quality of the original image.

Protecting facial images with PMask is beneficial in two aspects. First, individual users can use PMask to protect the photos they share online from being misused by privacy intruders to recognize their faces on images they collect, e.g., from a surveillance camera. Second, some privacy intruders build a face-based search engine [7] that can return the websites containing the victim's facial image, a serious threat beyond simply knowing the victim's identity (i.e., the name). Online content platforms, such as social networks, are the target of such crime due to their rich collections of their members' facial images. Hence, they can benefit from PMask by making the photos published on their websites unable to be matched to the correct person. This helps preserve their reputation by taking active methods to protect their members' digital footprints against misuse and abuse.

This paper makes the following original contributions. *First*, we develop a PMask framework to generate a privacy-preserving mask for enforcing privacy protection of each facial image prior to sharing it on the Internet. Our PMask generation process is iterative with continuous optimization of dual goals: (i) minimizing the amount of privacy noise injected while maximizing the effectiveness of hiding the true facial signature of a facial image, and (ii) minimizing the perception loss compared to the original facial image. The former ensures the image quality (i.e., no excessive noise), while the latter requires the protected image to be recognizable by humans. *Second*, to further improve the generalizability of PMask in hiding facial signatures by generating a robust privacy mask, we propose a principled approach using focal diversity-based ensemble learning to boost the robustness of PMask privacy protection. Instead of using a random ensemble, our focal diversity method can enhance the generalization performance of our privacy mask for more robust privacy protection. To the best of our knowledge, PMask is the first to utilize focal diversity for composing an ensemble of multiple FR models to identify and hide the unique facial signature of facial images by generating robust and generalizable privacy masks. We conduct extensive experiments to analyze PMask with eight FR models of different characteristics on FaceScrub [28], a widely-used FR benchmark. The results validate that the focal diversity-based privacy mask generation can deliver stronger protection against unknown FR models and outperforms the representative state-of-the-art methods with better user experience.

2 RELATED WORK

Existing approaches conduct pixel-level modification to facial images for privacy preservation [37]. They can be broadly classified into two categories. The first category uses FR models to craft small changes to perturb the original facial image. Fawkes [33] formulates an untargeted attack to push the facial image in the feature space away from the original location, while Face-off [12] uses a targeted objective. Several works also focus on the image quality with LowKey [13] using LPIPS [45] and TIP-IM [42] using MMD [10] to quantify and minimize the impact on human perception. PMask falls into this category and advances the above approaches in two aspects. First, we introduce a loss function to search for facial signatures with a convergence condition to just inject sufficient distortion without over-perturbing the face under protection. Second, all existing works manually select FR models to craft perturbations, but none studies the impact of selected FR models on robustness in privacy protection. PMask is the first to leverage the diversity-optimized ensemble teaming framework with a focal diversity-based ensemble selection method, showing its effectiveness in strengthening the privacy mask generation quality for stronger protection against unknown FR models. We argue that privacy-preserving facial masking should preserve two important data utilities: (1) The mask-transformed facial image should preserve the image quality, and (2) The identity of the mask-transformed facial image should remain recognizable by humans. Then, the masked facial image can be safely released in public. Recall Figure 1, PMask, Fawkes, and TIP-IM-protected gallery images fully preserve the second utility, although they differ in the first and the protection robustness.

In contrast to the pixel perturbation approaches, the second category uses conditional generative adversarial networks (GANs) [27] to synthesize a face similar to the original one [24, 25, 34]. Although the GAN-synthesized faces appear to be realistic facial images and satisfy our first utility above, it fails to meet the second utility. The owner can no longer recognize the person on the GAN-protected facial image, which appears to be a stranger and completely different from the original facial image. Hence, this category of work may not offer satisfactory user experience for human users.

3 PMASK OVERVIEW

DNN-based face recognition (FR) uses a feature extractor F trained to map a given facial image $\boldsymbol{x}$ to a high-dimensional vector $F(\boldsymbol{x})$ in the feature space. In addition to using a pretrained FR model F, the privacy intruder will also need to collect a set of gallery images $\mathcal{D}^G$ by web scraping [4] and utilizing publicly available face datasets. Then, the privacy intruder can use the pretrained FR model F to map each gallery image $\boldsymbol{x}^G \in \mathcal{D}^G$ to the feature vector $F(\boldsymbol{x}^G)$. Given a probe (query) image $\boldsymbol{x}^P$, the privacy intruder can use the same FR model F to map it to a feature vector $F(\boldsymbol{x}^P)$ and predict the identity of the "unknown" person by using the identity of the nearest gallery image, formally:

$$\mathcal{I}(\boldsymbol{x}^P; \mathcal{D}^G) = \mathcal{I}\left(\arg\min_{\boldsymbol{x}^G \in \mathcal{D}^G} \textsc{Dist}(F(\boldsymbol{x}^P), F(\boldsymbol{x}^G)); \mathcal{D}^G\right), \tag{1}$$

where $\mathcal{I}(\boldsymbol{x}; \mathcal{D}^G)$ denotes the identity of the image $\boldsymbol{x}$, which is known only for those gallery images in $\mathcal{D}^G$, and $\textsc{Dist}(\cdot, \cdot)$ is a distance function such as the Euclidean distance. The richer the gallery set the privacy intruder maintains, the higher the likelihood it will have one or more facial images of the person corresponding to the query. This can lead to a serious threat to user privacy.

To combat such a privacy threat, we introduce PMask, which develops two synergistic functional components of PMask loss optimization: The first component is to learn the search for the facial signature of a facial image and then the generation of a robust privacy mask to effectively hide the true identity from pretrained

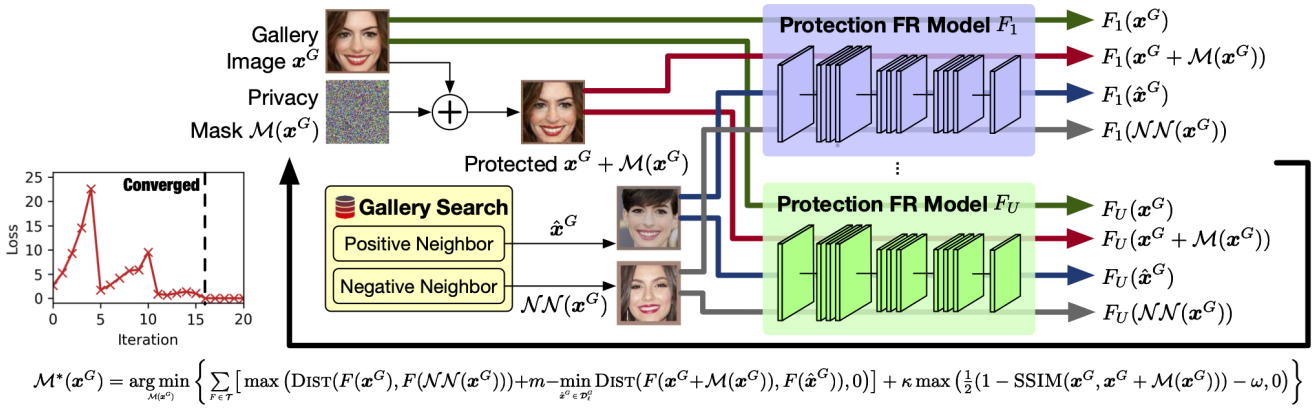


Figure 2: The iterative generation process of PMask to hide the unique signature from the given facial image.

FR model(s). The second component leverages the focal diversity-based ensemble selection method to find the most diverse team of FR models to boost the generalization performance of the first component and offer stronger protection with our privacy mask system. **Figure 2** gives an overview of the first component. Given a facial image to be protected, we leverage a team of carefully chosen FR models to iteratively generate the privacy mask until convergence, such that the facial signatures of the masked image will be hidden. When it is scraped by privacy intruders to build their gallery images, the masked image serves no purpose in identifying probe images of the person under protection.

Sometimes, the user may not have access to the photo that the adversary uses to query a face recognition model. For instance, a stalker can take a photograph of the victim and conduct FR on the face database constructed by web scraping (e.g., PimEyes [7]). The user cannot use a privacy tool to obtain privacy protection on this type of Web facial data. Instead, PMask shares the same threat model as the recent facial perturbation methods (e.g., Fawkes [33] and Lowkey [13]) and provides the facial images released online to be misused by intruders to invade their privacy. PMask improves over these existing methods with better protection effectiveness, image quality preservation, and protection speed. We also introduce a principled approach to selecting a high-quality team from a pool of pretrained FR models to boost protection against FR models.

In the following sections, we first introduce our optimization with dual objectives to find such a privacy mask (Section 4). Then, we present a principled approach to forming a team of FR models (Section 5) to strengthen the privacy mask against *un*known FR models, which is particularly important as a protector has no prior knowledge about which FR model will be used by a privacy intruder, be it FR algorithm, neural architecture, or training dataset.

4 PRIVACY MASK GENERATION

For each gallery image x^G of the person ℓ under protection, PMask generates a privacy mask $\mathcal{M}(x^G)$ to obfuscate its facial signatures. We can conceptualize the protection process to contaminate the set of gallery images $\mathcal{D}^G$ built by the privacy intruder through web scraping. Let $\mathcal{D}_\ell^G$ be the subset of gallery images belonging to the person ℓ under protection. PMask is to generate the protected set of gallery images $\tilde{D}^G$ formally as follows:

$$\begin{aligned} \tilde{D}^G &= \{x^G + \mathcal{M}(x^G) \mid x^G \in \mathcal{D}_\ell^G\} \cup [\mathcal{D} - \mathcal{D}_\ell^G] \\ \text{s.t. } & \mathcal{I}(x^P; \mathcal{D}^G) \neq \mathcal{I}(x^P; \tilde{\mathcal{D}}^G) \quad \forall x^P \in \mathcal{D}_\ell^P, \end{aligned} \tag{2}$$

where $\mathcal{D}_\ell^P$ is the set of probe images belonging to the person ℓ under protection. Intuitively, for each probe image x^P of the person ℓ, its identity recognized using the unprotected set of gallery images $\mathcal{D}^G$ should be different from the one using the PMask-protected set of gallery images $\tilde{D}^G$. Note that in the above formulation, we assume only one person is under protection for brevity, but it can be trivially extended to any or even all people.

To accomplish privacy protection, PMask needs to mask the gallery image so that it will not be matched as the nearest neighbor of a probe image of the same person. A straightforward solution is to perturb an image as much as possible such that the distance between the feature vector of the masked face and that of the unprotected face is maximized. However, the facial image will be significantly distorted, reducing the usability of PMask due to the loss of the two utility criteria of the PMask-transformed facial image: (i) it should preserve the image quality comparable to the original image in terms of perception similarity metrics; and (ii) its identity should remain recognizable by human, so the owner can safely publish the masked facial image online. PMask introduces dual optimizations to achieve privacy protection with minimal impact on image quality.

4.1 Reverse Triplet Optimization

We first introduce an identity-based reverse triplet loss for privacy protection in PMask. To motivate the reason behind the use of a reverse triplet loss, we first review the basic triple loss [32], a popular loss function for metric learning used in numerous applications, e.g., face recognition [32], object tracking [18], and cross-modal information retrieval [40]. It is defined by

$$\mathcal{L}_{\text{Tri}} = \max(\text{DIST}(F(x_a), F(x_+)) - \text{DIST}(F(x_a), F(x_-)) + m, 0), \tag{3}$$

where x_a is an anchor sample, x_+ is a positive sample with the same class as x_a, x_- is a negative sample with a different class than x_a, and m is a margin controlling how far should the negative sample be. Several sampling strategies have been introduced to form the triplets [19, 43].

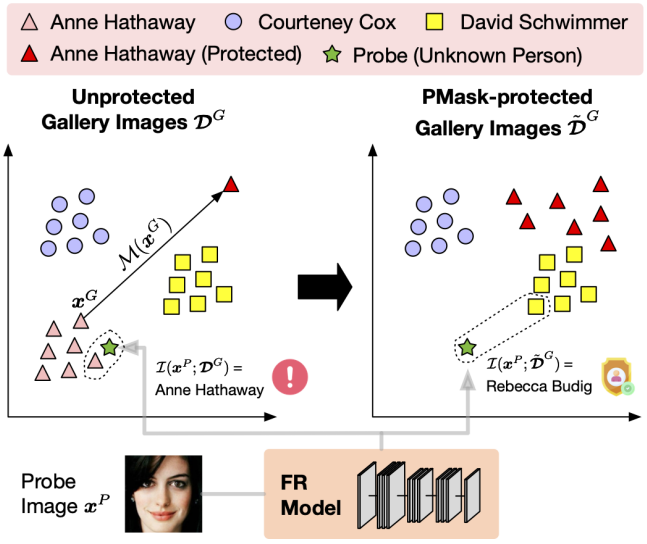


Figure 3: PMask pushes each gallery image of the person under protection (Anne Hathaway) away from the original position in the feature space (left). The FR based on the resultant gallery images (right) returns an image of Anne Hathaway to be Rebecca Budig.

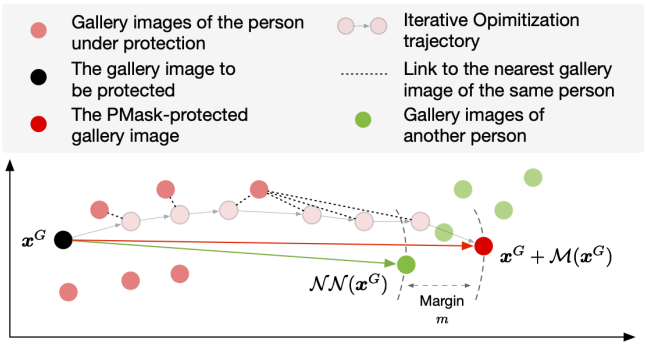


Figure 4: The reverse triplet loss progressively searches for the optimal location for the masked face to reside.

In the context of PMask, the direct adaptation of conventional triplet loss will not work. We create an identity-based reverse triplet loss function in order to correctly reflect the philosophy and the abstraction logic of conventional triplet loss in this privacy masking context. **Figure 3** gives an illustration. Considering the query image (the green star) as the anchor $\boldsymbol{x}_a$, we want our privacy masking learner to progressively learn a feature extractor to bring the embeddings of those positive images in the gallery set of the same class (person) as the anchor $\boldsymbol{x}_a$, i.e., the set of $\boldsymbol{x}_+$'s (highlighted in pink triangles) to be farther away from $\boldsymbol{x}_a$ than the distance of the anchor $\boldsymbol{x}_a$ to some negative images (i.e., $\boldsymbol{x}_-$) from a different class.

The term "*reverse*" has multiple meanings. First, given an image from the gallery set and its actual personal label (the full name of a person), PMask iteratively learns to generate a privacy mask that can effectively prevent the masked image from being mapped to the name labeled on the original input image. The mask generation process first freezes the FR models used for learning the mask and randomly adds a small amount of noise. Then, it utilizes PMask's loss optimizer to continuously learn the right amount and pixel locations for fine-tuning the noise to mask the facial image to be protected. Hence, the learning task and objective are different from the conventional procedure. Second, to learn the privacy mask of a given gallery image, PMask's loss optimization will have two components: one for making sure the perturbation effectively hides the protected gallery image in the crowd with a new identity, and the other is to ensure the masked image has a similar visual perception to the original one in both digital measures and the view of human. The former ensures privacy is preserved after masking, and the latter ensures the desired visual utility is preserved on the masked image under the protection of PMask. We use the reverse triplet loss because we aim to "maximize" the distance between the samples of the same class in the gallery set and the anchor image, i.e., the input image to be masked by PMask. At the same time, we aim to "minimize" the distance between this anchor image and some samples of different classes. Hence, we call our first loss component the reverse triplet loss. **Figure 4** provides another perspective for the illustration of our reversed triplet loss. The masked image $\boldsymbol{x}^G + \mathcal{M}(\boldsymbol{x}^G)$ should be farther away from the embedding location of the original unprotected one $\boldsymbol{x}^G$ than the nearest gallery image $\mathcal{NN}(\boldsymbol{x}^G)$ of a different class (person) by a margin m. The iterative optimization to ensure the protected image resulting from adding the privacy mask learned iteratively (i.e., $\boldsymbol{x}^G + \mathcal{M}(\boldsymbol{x}^G)$) is far away from *any* gallery image of the same class as the original input image. Given that PMask uses a protection ensemble of FR models $\mathcal{T}$ to generate the privacy mask for each input facial image, the following reverse triplet loss function will protect the given image $\boldsymbol{x}^G$ of the class labeled as person ℓ:

$$\mathcal{L}_{\text{RevTriplet}} = \sum_{F \in \mathcal{T}} \Big[\max(\textsc{Dist}(F(\boldsymbol{x}^G), F(\mathcal{NN}(\boldsymbol{x}^G))) + m - \min_{\hat{\boldsymbol{x}}^G \in \mathcal{D}^G_\ell} \textsc{Dist}(F(\boldsymbol{x}^G + \mathcal{M}(\boldsymbol{x}^G)), F(\hat{\boldsymbol{x}}^G)), 0) \Big], \tag{4}$$

where m is a hyperparameter controlling the margin. To provide a more intuitive setting, we redefine it to be a fraction δ of the distance between the unprotected image and its nearest gallery image of a different person:

$$m = \delta \cdot \textsc{Dist}(F(\boldsymbol{x}^G), F(\mathcal{NN}(\boldsymbol{x}^G))). \tag{5}$$

A nice property in Equation 4 is that the loss becomes zero when the first term is non-positive. Then, no further perturbation is required. As shown in Section 6, the number of required iterations depends on the hardness of the image to be protected and can be determined on the fly by PMask.

4.2 Imperceptibility Optimization

Even though the reverse triplet loss can generate privacy masks protecting users, it is not attractive if the masked image is seriously distorted. Hence, we dedicate the second optimization to the imperceptibility of privacy masks. The structural similarity (SSIM) index [36] is a popular image quality measure comparing two images. Instead of comparing images by per-pixel differences such

as the mean square error, SSIM focuses on the structure information on the image, which aligns with the human visual perception system and is defined as follows:

$$\text{SSIM}(\boldsymbol{x}^G, \tilde{\boldsymbol{x}}^G) = \frac{2\mu_{\boldsymbol{x}^G}\mu_{\tilde{\boldsymbol{x}}^G}}{\mu_{\boldsymbol{x}^G}^2 + \mu_{\tilde{\boldsymbol{x}}^G}^2} \frac{2\sigma_{\boldsymbol{x}^G}\sigma_{\tilde{\boldsymbol{x}}^G}}{\sigma_{\boldsymbol{x}^G}^2 + \sigma_{\tilde{\boldsymbol{x}}^G}^2} \frac{\sigma_{\boldsymbol{x}^G\tilde{\boldsymbol{x}}^G}}{\sigma_{\boldsymbol{x}^G}\sigma_{\tilde{\boldsymbol{x}}^G}}, \tag{6}$$

where $\mu_{\boldsymbol{x}}$ is the mean luminance of an image $\boldsymbol{x}$, $\sigma_{\boldsymbol{x}}$ is the standard deviation, and $\sigma_{\boldsymbol{x}\tilde{\boldsymbol{x}}}$ is the covariance between two images. It ranges from -1 to $+1$, where a higher value means two images are perceptually more similar. To control the privacy mask to generate a protected image perceptually similar to the unprotected one, we define the imperceptibility loss as follows:

$$\mathcal{L}_{\text{Impercept}} = \max(\frac{1}{2}(1 - \text{SSIM}(\boldsymbol{x}^G, \boldsymbol{x}^G + \mathcal{M}(\boldsymbol{x}^G))) - \omega, 0), \tag{7}$$

where ω controls the SSIM degradation. It ensures that the image quality, measured in SSIM, does not fall below ω. We chose SSIM because it is a popular image quality measure that aligns well with the human visual perception system. There are several reference systems available to help select the desired SSIM (e.g., ">0.99" is excellent, and "0.95-0.99" is good [47]). We first rescale SSIM from the range of [-1, 1] to [0, 1]. Then, if the image quality at the current iteration falls below the threshold, a non-zero loss will be produced, which will require optimization to reduce perturbations. Similar to the reverse triplet loss in Equation 4, the above imperceptibility loss also has a convergence condition (i.e., SSIM $> 1 - 2\omega$).

Based on the above, we can find the privacy mask for the image $\boldsymbol{x}^G$ of person ℓ as follows:

$$\mathcal{M}^*(\boldsymbol{x}^G) = \underset{\mathcal{M}(\boldsymbol{x}^G)}{\arg\min} \; [\mathcal{L}_{\text{RevTriplet}} + \kappa \mathcal{L}_{\text{Impercept}}], \tag{8}$$

where κ is a regularization hyperparameter balancing the reverse triplet loss and the imperceptibility loss. We use a dynamic schedule to define κ with an initial value of 1.0, which will be doubled or halved when the amount of perturbation is too low or high [33]. Overall, the optimization is complete when the protection objective defined in Equation 4 is achieved, and the degradation in image quality is within the budget in Equation 7.

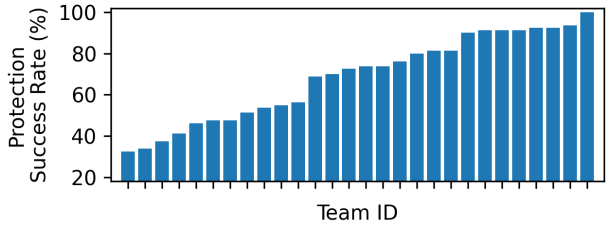
5 FOCAL DIVERSITY TEAMING

We have introduced the privacy mask generation using a given team of FR models. The next question to be answered is how to find a team offering strong privacy protection. A large number of models are publicly available. They can be directly applied to production systems because FR models can extract features for faces unknown during the training process. This allows one to create a collection of FR models easily. **Table 1** shows an example collection of eight FR models. They use ArcFace [17], the state-of-the-art FR algorithm, requiring an input resolution of 112×112 and encoding a given facial image into a 512-dimensional feature vector. They differ in terms of neural architectures (i.e., EfficientNet [35] or ResNet [22] in the 2nd column) and training datasets (i.e., MS-Celeb-1M [21], Glint360K [9], VGG-Face2 [11], WebFace600K [46] in the 3rd column). All models achieve a competitive FR accuracy on the FaceScrub [28] dataset for testing (i.e., 4th column).

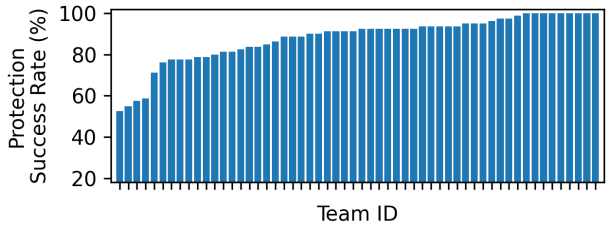
While one could use all available models for protection, prior works have shown a carefully chosen subset of models can lead to better generalization performance in image classification [38, 39] and object detection [15]. To demonstrate the necessity of teaming

Table 1: The collection of face recognition (FR) models.

Model ID	Architecture	Training Dataset	Accuracy
F_1	EfficientNet	MS-Celeb-1M	94.94%
F_2	ResNet50	MS-Celeb-1M	89.25%
F_3	ResNet50	Glint360K	96.68%
F_4	ResNet50	VGG-Face2	94.30%
F_5	ResNet50	WebFace600K	96.72%
F_6	ResNet18	MS-Celeb-1M	82.64%
F_7	ResNet34	MS-Celeb-1M	84.49%
F_8	ResNet100	MS-Celeb-1M	91.85%



(a) Two-member Teams: $\binom{8}{2} = 28$ options



(b) Three-member Teams: $\binom{8}{3} = 56$ options

Figure 5: The effectiveness of protection teams can vary. Finding a high-quality combination is crucial.

in privacy protection, we enumerate all 28 teams of size two from the collection. For each team, we measure the protection success rate against a privacy intruder. **Figure 5a** shows that a poor choice can lead to weak protection. Only 32.50% of the testing images are incorrectly identified by the privacy intruder. In contrast, a good choice can offer substantial protection, meaning the privacy intruder will not be able to identify the correct name of an unknown face. A similar observation can be found for three-member teams (56 teams in total) in **Figure 5b**. Such a divergence in protection effectiveness confirms the need for a principled approach to identify a high-quality protection team. However, evaluating each possible protection team with a validation set is impractical because of the time complexity, even though PMask is already more efficient than existing approaches (details are provided in Section 6.1).

We introduce a diversity-driven approach to conduct efficient and effective teaming. The intuition is to form a team of FR models making diverse mistakes, implying their divergence in the decision-making process. If the privacy mask can protect against such a

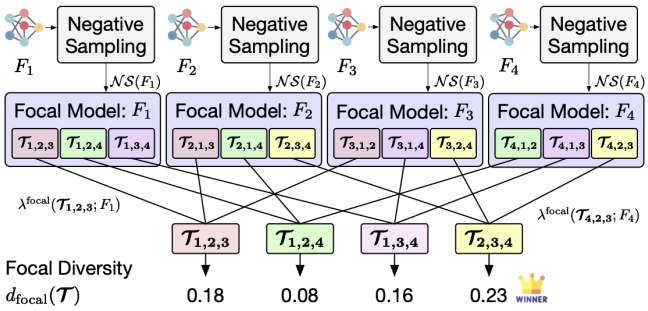


Figure 6: The workflow of focal diversity to efficiently quantify the quality of a protection team of FR models. Note that the protection team $\mathcal{T}_{i,j,k}$ means $\{F_i, F_j, F_k\}$.

diverse team, it is more likely to transfer to disable other FR models. Our design is based on the focal diversity framework originally proposed for image classification systems [38]. **Figure 6** provides an overview of the workflow considering a collection of four models comparing teams of size three. Concretely, given a collection of N FR models $\{F_1, ..., F_N\}$, we first identify the negative samples, $\mathcal{NS}(F)$, for each FR model F by locating validation images that F fails to recognize their true identity and remove those negative samples that are negative w.r.t. all FR models (i.e., $\mathcal{NS}(F_1) \cap ... \cap \mathcal{NS}(F_N)$). To rank protection teams of size S, we enumerate all $\binom{N}{S}$ combinations, denoted by $\text{ProtTeams}_{\text{size}=S}$. For each team $\mathcal{T} \in \text{ProtTeams}_{\text{size}=S}$, we consider each member to be the focal model F_{focal} and use its negative samples $\mathcal{NS}(F_{\text{focal}})$ to statistically estimate the level of negative correlation between F_{focal} and the remaining models in $\mathcal{T}$. The focal negative correlation of team $\mathcal{T}$ w.r.t. the focal model F_{focal}, denoted by $\lambda^{\text{focal}}(\mathcal{T}; F_{\text{focal}})$, is computed by measuring the degree of disagreements using a generalized non-pairwise measure [29]: let Y denote a random variable representing the proportion of models (i.e., i out of S) that fail to recognize a random negative sample $\boldsymbol{x} \in \mathcal{NS}(F_{\text{focal}})$. With the propability $Y = \frac{i}{S}$ denoted as p_i, the focal generalized negative correlation can be computed as

$$\lambda^{\text{focal}}(\mathcal{T}; F_{\text{focal}}) = \frac{\sum_{i=1}^{S} \frac{i}{S} p_i}{\sum_{i=1}^{S} \frac{i(i-1)}{S(S-1)} p_i} \tag{9}$$

with a range of $[0, 1]$ having the minimum correlation of 0 when the failure of one member is accompanied by the correct recognition by the other. The same procedure is repeated by considering each team member as the focal model, and the focal diversity of the protection team $\mathcal{T}$ is finalized as follows:

$$d_{\text{focal}}(\mathcal{T}) = \frac{1}{S} \sum_{F_{\text{focal}} \in \mathcal{T}} [1 - \lambda^{\text{focal}}(\mathcal{T}; F_{\text{focal}})]. \tag{10}$$

With focal diversity, users can select a high-quality team within their computing budget according to the high diversity score.

Figure 7 gives an end-to-end overview of PMask. We provide a service for users to perturb their images with PMask before sharing such that even if they are scraped, no meaningful feature vectors can be extracted. Given an image, PMask first uses an existing face detection algorithm (e.g., MTCNN [44]) to detect and crop the segments containing a face. Then, for each facial image, we use the high-diversity team with our optimization algorithm with dual objectives to search for a privacy mask and apply it to perturb the corresponding region in the original image. Once all faces on the image have been masked, the PMask-protected image can be shared by the user on, e.g., social media platforms.

6 EXPERIMENTAL EVALUATION

We conduct extensive experiments on the representative benchmark dataset for face recognition, FaceScrub [28], to analyze the effectiveness of PMask. It consists of 50, 924 facial images of 530 celebrities. PMask inherits the one-shot learning nature of FR and is applicable to any face dataset. Since commercial FR APIs like Azure now require manual approval to use their services to avoid misuse, we expect that the privacy intruder will opt for open-source FR models as they are readily available. Hence, our experiments focus on protecting FR against pretrained models that are publicly available online. We study PMask using the collection of FR models reported in Table 1. Note that none of the models uses FaceScrub (i.e., the testing set) for training, which is necessary to provide fair evaluation. In our experiments, we consider ten randomly chosen celebrities for protection analysis. The probe set consists of 100 facial images, 10 from each selected celebrity. Their remaining facial images, together with all images of 520 other non-chosen celebrities, form the gallery set, which consists of 50, 824 images.

We compare PMask with two representative approaches, Fawkes [33] and TIP-IM [42], following the default settings in their open-source

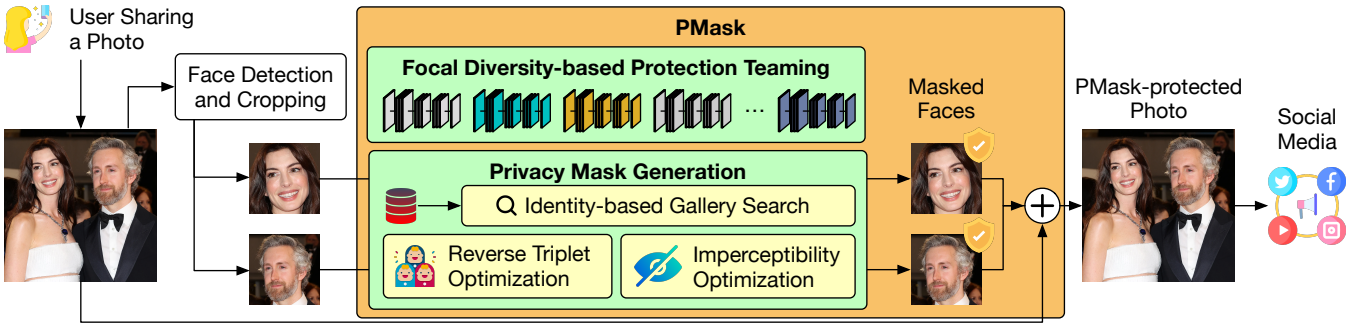


Figure 7: The overview of PMask workflow to apply a privacy mask for each face on the given image.

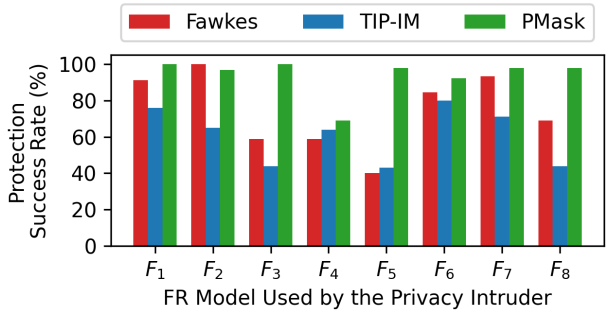


Figure 8: The comparison of Fawkes, TIP-IM, and PMask against the privacy intruder using eight different FR models.

repository. For fair comparisons, the perturbation budget ω is set to 0.017 for all approaches, which is the same as the one used in Fawkes. By default, we set δ in Equation 5 to be 1.0. This hyper-parameter controls the trade-off between image quality, protection time, and effectiveness. We will provide an analysis of it in Section 6.4. TIP-IM uses an ArcFace [17] model with a ResNet50 architecture, Fawkes uses the two-member team (F_1, F_2) in Table 1, and the default setting for PMask is (F_1, F_3), the most diverse team identified by our teaming method. The source code of PMask is publicly available at https://github.com/git-disl/PMask. All measurements are recorded on NVIDIA RTX 2080 SUPER GPU, Intel i7-9700K (3.60GHz) CPU, and 32 GB RAM on Ubuntu 18.04.

6.1 Privacy Protection Analysis

In this subsection, we analyze PMask from three perspectives: (i) protection effectiveness, (ii) image quality, and (iii) time cost.

Better Protection. To evaluate the protection effectiveness against a privacy intruder using an FR model F, we define the protection success rate (PSR) in percentage to be (100 − FR Accuracy)%. Intuitively, our goal is to maximize the PSR such that the privacy intruder will not be able to identify a person's true identity given the facial image. **Figure 8** reports the PSR against each of the eight FR models. We make several interesting observations. First, F_1 is the model used by both PMask and Fawkes to generate protection masks, but only PMask can provide perfect protection with a PSR of 100%, meaning that Fawkes may not provide strong protection even if it knows which FR model the privacy intruder will use. Second, F_2 is the second model used by Fawkes, which reaches a PSR of 100%. Even though PMask does not have access to this model during the protection process, it still achieves a high PSR of 97% in this black-box protection setting. Third, TIP-IM offers a certain level of protection, but it can only outperform Fawkes when the privacy intruder uses F_4 or F_5 as the FR model. In contrast, PMask reaches a much better PSR than both Fawkes and TIP-IM when the privacy intruder uses F_4, F_5, F_6, F_7, or F_8, which are all unknown to protection mechanisms.

We further report the PSR of three celebrities in **Table 2** to show that PMask is not just, on average, effective but can be consistently helpful for different celebrities. For instance, when the privacy intruder uses F_5, even under the protection of Fawkes or TIP-IM, all probe images of David Schwimmer can still be perfectly identified (i.e., a PSR of 0%). In contrast, under PMask's protection, the same FR model can only identify 10% of the probe images (i.e., a PSR of 90%). We can make similar observations across different celebrities and FR models being used by the intruder.

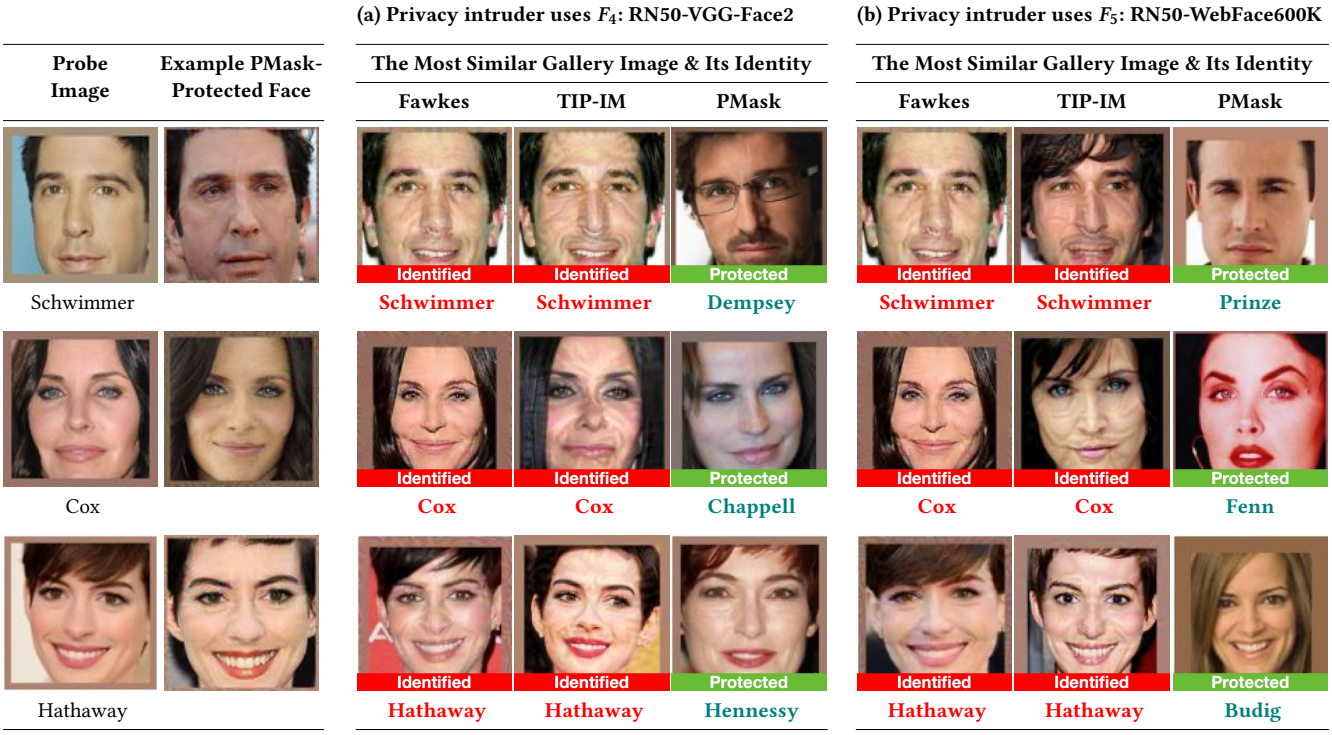
Table 2: PMask offers the most consistently effective protection, while other approaches vary across various celebrities and FR models used by the privacy intruder.

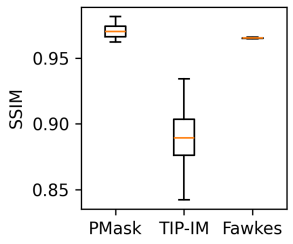
Name	Method	Protection Success Rate (%)							
		F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
Unprotected		0	0	0	0	0	0	0	0
David Schwimmer	Fawkes	**100**	**100**	20	50	0	70	90	30
	TIP-IM	0	0	0	30	0	20	20	0
	PMask	**100**	90	**100**	**80**	**90**	**90**	**100**	**100**
Courteney Cox	Fawkes	**100**	**100**	60	70	60	70	**80**	50
	TIP-IM	**100**	90	90	90	80	60	**80**	80
	PMask	**100**	90	**100**	**90**	**100**	**80**	**80**	**100**
Anne Hathaway	Fawkes	**100**	**100**	20	**80**	30	**100**	70	30
	TIP-IM	90	90	30	70	10	90	90	30
	PMask	**100**	90	**100**	**80**	**100**	**100**	**100**	**80**

For each of the three celebrities in Table 2, we select one example probe image and use (a) F_4 and (b) F_5 to find the most similar image from the gallery set protected by Fawkes, TIP-IM, and PMask in **Table 3**. Both FR models are unknown to all protection mechanisms. Under the protection by Fawkes (3rd and 6th columns) and TIP-IM (4th and 7th columns), the most similar gallery image belongs to the same person as the corresponding probe image, meaning the protection fails. Yet, PMask (5th and 8th columns) can deceive both unknown FR models to malfunction. For instance, David Schwimmer's face (1st example) is matched to Patrick Dempsey's by F_4 in (a) and to Freddie Prinze's by F_5 in (b). Similarly, both Courteney Cox's (2nd example) and Anne Hathaway's (3rd example) faces are mismatched to the wrong person under the protection of PMask.

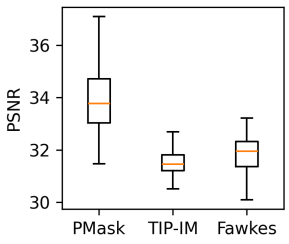
Better Image Quality. No one would like to share a selfie with a significantly distorted face. Hence, the privacy-protected image to be shared online should be of high image quality and look similar to the original one. We use two metrics most commonly used across various fields to quantify image quality. **Figure 9** gives the distributions of SSIM [36] and PSNR [23] of images protected by Fawkes, TIP-IM, and PMask, demonstrating the overall image quality under protection. A higher value means better quality. We make three observations. First, PMask and Fawkes generate protected images of much higher SSIM than TIP-IM. The degraded image quality by TIP-IM can be observed in the 4th and the 7th columns in Table 3, where the protected faces are seriously distorted. Second, PMask-protected images tend to have a higher SSIM than Fawkes, even if both explicitly minimize the SSIM degradation. The improvement made by PMask can be attributed to the convergence condition in Equation 4, where PMask stops perturbing the image when the condition is met, while Fawkes continues to do so until a predefined number of iterations is executed. Third, even though none of the

Table 3: The most similar gallery image found by two FR models given the probe image (1st column). Both FR models used by the privacy intruder are unknown during any protection process. They can still find the most similar gallery image with the correct identity under Fawkes' and TIP-IM's protection, but PMask leads to the wrong one (i.e., successful protection).



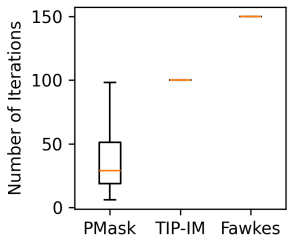


(a) SSIM (Higher is better)

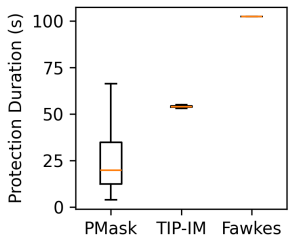


(b) PSNR (Higher is better)

Figure 9: The distributions of SSIM and PSNR over 100 gallery images protected by PMask, TIP-IM, and Fawkes. PMask-protected gallery images reach higher SSIM and PSNR than both Fawkes and TIP-IM.



(a) Required Iterations

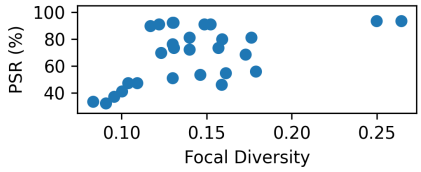


(b) Duration per Face

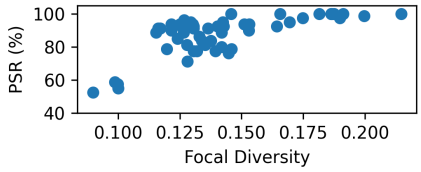
Figure 10: The distributions of (a) the required iterations and (b) the protection duration per face over 100 gallery images protected by PMask, TIP-IM, and Fawkes. TIP-IM and Fawkes require a fixed number of iterations (i.e., 100 and 150, respectively). In contrast, the number of iterations needed by PMask varies across facial images because PMask can finish the iterative process sooner.

protection mechanisms uses PSNR to optimize the imperceptibility of perturbations, PMask generates protected images with a better (higher) PSNR, while Fawkes and TIP-IM tend to produce images with a similar quality in terms of PSNR.

Better Time Cost. A lengthy protection duration can prevent users from sharing their, e.g., social media posts immediately and reduce the attractiveness of employing privacy protection. **Figure 10a** compares the required iterations per face between PMask, TIP-IM, and Fawkes. The number of required iterations for PMask varies across faces and can be determined automatically by PMask based on the hardness to satisfy the condition defined in Equation 8. Once convergence is reached, PMask terminates. However, TIP-IM and Fawkes require a fixed number of iterations (i.e., 100 and 150, respectively). The number of required iterations directly affects the protection duration per face, as reported in **Figure 10b**. Most faces



(a) *Two*-member Teams



(b) *Three*-member Teams

Figure 11: Ranking teams with diversity.

Table 4: The most diverse and the least diverse teams with (a) two or (b) three members identified by our teaming with their PSR against different FR models used by the privacy intruder.

Protection Team	Protection Success Rate (%)									
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	Mean	Std
Unprotected	0	0	0	0	0	0	0	0	0.00	0.00
(a) *Two*-member Teams										
Most Diverse: F_1, F_3	100	90	100	80	100	100	100	80	**93.75**	**9.16**
Least Diverse: F_6, F_7	0	50	0	10	0	100	100	10	33.75	44.06
(b) *Three*-member Teams										
Most Diverse: F_1, F_3, F_5	100	100	100	100	100	100	100	100	**100.00**	**0.00**
Least Diverse: F_2, F_6, F_7	0	100	60	20	0	100	100	40	52.50	44.00

can be protected within 50 seconds by PMask, but Fawkes requires over 100 seconds, which is 2× slower than ours. This time cost is measured per face. In real life, dozens of people can appear in the same image, and the advantage of PMask is much more significant.

6.2 Focal Diversity-driven Protection Teaming

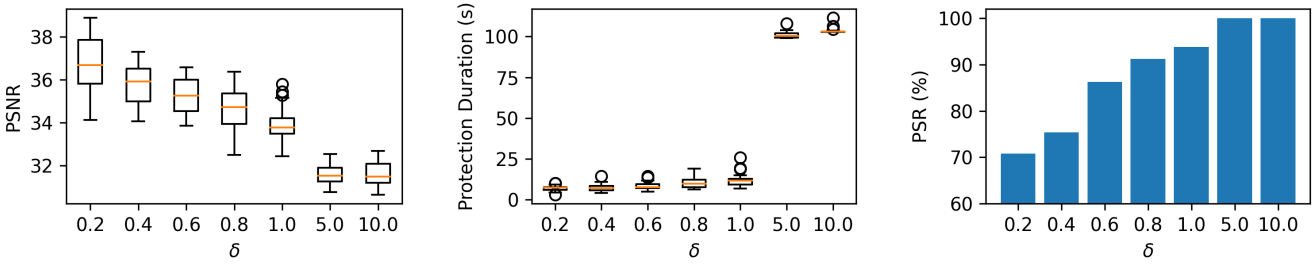
In this subsection, we analyze our focal diversity to understand how well it can identify protection teams that can lead to stronger protection when used to generate privacy masks. Given the pool of eight models in Table 1, we enumerate all possible teaming combinations with two and three members. We show the focal diversity of each protection combination and its PSR in **Figure 11a** and **Figure 11b**, respectively. We use the PSR for the celebrity Anne Hathaway in the following experiments due to a similar trend in other celebrities. According to the scatter plots, our focal diversity positively correlates with the PSR, meaning that choosing a team (combination) with high focal diversity can generate privacy masks with stronger protection. This can be observed in **Table 4**, where we provide the detailed PSR against each FR model using the most diverse and the least diverse protection teams identified by our focal diversity for a two-member setting (Table 4a) and a three-member setting (Table 4b). An interesting observation is that a team of models having different neural architectures is not always the top priority [14]. Focusing on the three-member case as an example, the most diverse team is (F_1, F_3, F_5), which leads to perfect protection against all FR models. The FR models F_1 and F_3 are of different neural architectures (EfficientNet and ResNet50). While there are other FR models with a different architecture than these two members (e.g., ResNet18, ResNet34, or ResNet100), our focal diversity teaming selects F_5 as the third member, which is also a ResNet50 as F_3 but trained on a different dataset. For the least diverse team (F_2, F_6, F_7), all members are trained on the same dataset (i.e., MS-Celeb-1M) but with different depths in ResNet. Such a team can only lead to perfect protection when the privacy intruder uses the same FR model (i.e., F_2, F_6, or F_7). Otherwise, the protection is weak, and the average PSR is only 52.50%. The teaming method can be easily extended to consider a wide range of backbones in the FR model collection. With more backbones included, a critical optimization is to select a subset of FR models to form an ensemble team instead of using all backbone models. The results show that selecting a team of highly diverse FR models is crucial to generating privacy masks. A higher diversity can lead to protected faces that can effectively transfer to deceive other unknown FR models utilized by the privacy intruder.

Table 5: An ablation study of PMask with different variants.

	Protection Success Rate (%)									
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	Mean	Std
Unprotected	0	0	0	0	0	0	0	0	0	0
Fawkes	**100**	**100**	20	80	30	**100**	70	30	52.50	44.00
(a) Rev. Triplet Loss	**100**	**100**	40	70	50	**100**	80	60	75.00	23.90
(b) Focal Div. Team	**100**	90	**100**	70	30	**100**	80	50	77.50	26.05
(c) PMask: (a) + (b)	**100**	90	**100**	**80**	**100**	**100**	**100**	**80**	**93.75**	**9.16**

6.3 Ablation Studies

The synergy of the reverse triplet loss and the focal diversity protection teaming provides strong protection effectiveness. We demonstrate the necessity of both modules by conducting an ablation study in this subsection. **Table 5** uses Fawkes as the baseline and compares it with three variants of PMask: (a) our reverse triplet loss with Fawkes's team (F_1, F_2), (b) Fawkes's loss with our diverse two-member team (F_1, F_3), and (c) our reverse triplet loss with our diverse two-member team (F_1, F_3), which is the complete PMask. Using either (a) or (b) can lead to much better protection, boosting the mean PSR across all FR models from 52.50% by Fawkes to 75.00% by (a) or 77.50% by (b). The standard deviations also drop drastically by almost half (the 11th column), meaning the protection against different unknown FR models is more stable. When both modules are enabled, the mean PSR further increases to 93.75%, which is significantly better than employing only one module and Fawkes, and the standard deviation drops to only 9.16%. For instance, the



(a) Image Quality **(b) Protection Time** **(c) Protection Effectiveness**

Figure 12: Hyperparameter analysis on δ to control the trade-off between image quality, protection time, and effectiveness.

protection against F_5 has only a success rate of 50% by (a). Using the complete PMask can enhance the protection to 100%.

6.4 Trade-off Analysis

The main hyperparameter in PMask is the δ in Equation 5, controlling how far the margin should be considered sufficient to converge. This subsection analyzes its impact on image quality, time cost, and protection effectiveness. **Figure 12** reports the trade-off between these three factors in different settings of δ. Setting a small δ (e.g., 0.20) means only a slight overshoot in the feature space is already sufficient to consider reaching convergence. Since it implies that the protection process can terminate sooner, we can expect the image quality to be better (Figure 12a) and the protection time to be shorter (Figure 12b). However, it may lead to insufficient perturbations and weaker protection (Figure 12c). Comparatively, a large δ (e.g., 10) can provide strong protection, but the image quality drops, and the protection time becomes longer. In practice, one could use a small number of validation images with a grid search to find the appropriate setting based on the application scenario. Several settings can be provided to users such that they can select the one that suits their needs. For instance, a person with a more serious privacy concern can choose a larger δ for better protection.

6.5 PMask Against Adaptive Intruders

A privacy intruder with advanced knowledge may introduce additional mechanisms to disable the countermeasure. We study the effectiveness of PMask under such adaptive intruders by considering four *adaptive attacks* aiming to wash out PMask. **Table 6** summarizes the results under (a) JPEG-compression attack [16], (b) mean-filter attack, (c) gaussian-filter attack, and (d) median-filter attack [41]. For JPEG compression, we use a low quality of 70 to launch an aggressive compression. For spatial filtering-based methods, the window size is set to 3×3. These are popular methods studied in the adversarial robustness domain [20]. We observe that PMask degrades when adaptive attacks are launched, but the impact is limited. Our perception loss, which requires the preservation of the visual quality of protected faces, makes the patterns introduced by PMask smoother and, hence, less sensitive to input transformation-based attacks. PMask is more resilient to the JPEG-compression attack with a drop in PSR of 7.50% even under a highly aggressive compression. Comparatively, it is more sensitive to spatial filtering-based approaches. Still, the PSR against different FR models can be maintained at least 72.50%. While the adaptive intruder may lower the quality parameter in JPEG compression or increase the window size in spatial filtering to further remove the patterns introduced by PMask, the operation also removes the salient features of faces that are necessary for face recognition. In other words, the privacy intruder needs to take the risk of lowering the FR accuracy due to low-quality facial images.

Table 6: The protection success rate of PMask against an adaptive intruder using different strategies to remove the patterns introduced by PMask to protect the images.

	Protection Success Rate (%)									
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	Mean	Std
Non-adaptive Baseline	100	90	100	80	100	100	100	80	93.75	9.16
(a) JPEG-compr. Attack	90	80	100	80	90	80	90	80	86.25	7.44
(b) Mean-Fil. Attack	70	70	80	80	70	70	80	60	72.50	7.07
(c) Gauss-Fil. Attack	90	80	90	70	80	80	80	70	80.00	7.56
(d) Med-Fil. Attack	90	90	80	70	80	90	90	70	82.50	8.86

Alternatively, the privacy intruder may attempt to retrain the FR model with PMask-protected images. The main difficulty is that the adversary needs to first identify those PMask-protected images from the large pool of photos scraped on the Internet. Then, this adversary can train the FR model on those PMask-protected images, which may lead to poor FR accuracy on unprotected images. Furthermore, separating protected and unprotected images may not be feasible due to its high cost and the imperceptibility of PMask-injected perturbations.

Under the scenario in which Web users did not protect all their facial images by PMask and the adversary may have some facial images of a user, say, Alice, PMask is still useful in protecting the user's privacy, especially PMask helps to reduce their digital footprints. Consider the threat scenario in which face search engines (built by the intruder or companies like PimEyes [7]) continuously collect facial images by web scraping. Given a facial image of Alice, those companies can retrieve all web pages containing images of

Alice's face. This is a serious privacy issue as Alice's digital footprint will be disclosed to, e.g., her stalker. Alice can use PMask to protect her new facial images before posting them online. Then, those new web pages containing Alice's protected facial images will not be matched as part of Alice's digital footprint.

7 CONCLUSION

We have presented PMask for anti-facial recognition through privacy protection masks. First, we develop a privacy mask generation process to learn the unique signatures of facial images to be protected and hide them from unauthorized FR models. It is accomplished by iterative optimization with dual goals: (i) maximizing the effectiveness of privacy protection and (ii) minimizing the perception loss compared to the original facial image. Second, we further improve the generalizability of PMask with a principled approach using focal diversity-based ensemble learning to enhance the protection effectiveness against unknown FR models. Our future work includes further speeding up the protection and exploring the authorization of FR models owned by trusted entities.

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] 2019. Brazilian retailer quizzed over facial recognition tech. https://www.zdnet.com/article/brazilian-retailer-quizzed-over-facial-recognition-tech/.
[2] 2021. 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/.
[3] 2021. The Secretive Company That Might End Privacy as We Know It. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.
[4] 2023. Clearview AI. https://www.clearview.ai/.
[5] 2023. Facial recognition and identity risk. https://www.equifax.co.uk/resources/identity-protection/facial-recognition-and-identity-risk.html.
[6] 2023. InsightFace: 2D and 3D Face Analysis Project. https://github.com/deepinsight/insightface.
[7] 2023. PimEyes. https://pimeyes.com/.
[8] Le An, Mehran Kafai, and Bir Bhanu. 2013. Dynamic Bayesian network for unconstrained face recognition in surveillance camera networks. *IEEE Journal on emerging and selected topics in circuits and systems* 3, 2 (2013), 155–164.
[9] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. 2021. Partial fc: Training 10 million identities on a single machine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1445–1449.
[10] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. 2006. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, 14 (2006), e49–e57.
[11] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.
[12] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. 2021. Face-off: Adversarial face obfuscation. *Proceedings on Privacy Enhancing Technologies* (2021).
[13] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. 2022. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*.
[14] Ka-Ho Chow and Ling Liu. 2021. Robust Object Detection Fusion Against Deception. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery.
[15] Ka-Ho Chow and Ling Liu. 2022. Boosting Object Detection Ensembles with Error Diversity. In *2022 IEEE International Conference on Data Mining (ICDM)*. IEEE.
[16] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. 2018. 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*. 196–204.
[17] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.
[18] Xingping Dong and Jianbing Shen. 2018. Triplet loss in siamese network for object tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 459–474.
[19] Weifeng Ge. 2018. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 269–285.
[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. 2017. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117* (2017).
[21] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic imaging* 2016, 11 (2016), 1–6.
[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[23] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.
[24] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. 2022. 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*. 15014–15023.
[25] Tao Li and Lei Lin. 2019. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
[26] Lizheng Liu, Bo Zhou, Zhuo Zou, Shih-Ching Yeh, and Lirong Zheng. 2018. 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)*. IEEE, 1–4.
[27] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
[28] Hong-Wei Ng and Stefan Winkler. 2014. A data-driven approach to cleaning large face datasets. In *2014 IEEE international conference on image processing (ICIP)*. IEEE, 343–347.
[29] Derek Partridge and Wojtek Krzanowski. 1997. Software diversity: practical statistics for its measurement and exploitation. *Information and software technology* 39, 10 (1997), 707–717.
[30] Nicolas Pinto, Zak Stone, Todd Zickler, and David Cox. 2011. Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *CVPR 2011 WORKSHOPS*. IEEE, 35–42.
[31] Mrutyunjaya Sahani, Chiranjiv Nanda, Abhijeet Kumar Sahu, and Biswajeet Pattnaik. 2015. 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]*. IEEE, 1–6.
[32] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.
[33] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20)*. 1589–1604.
[34] Qianru Sun, Ayush Tewari, Weipeng Xu, Mario Fritz, Christian Theobalt, and Bernt Schiele. 2018. A hybrid model for identity obfuscation by face replacement. In *Proceedings of the European conference on computer vision (ECCV)*. 553–569.
[35] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.
[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
[37] E. Wenger, S. Shan, H. Zheng, and B. Y. Zhao. 2023. SoK: Anti-Facial Recognition Technology. In *2023 IEEE Symposium on Security and Privacy (SP)*. 134–151.
[38] Yanzhao Wu and Ling Liu. 2021. Boosting Deep Ensemble Performance with Hierarchical Pruning. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1433–1438.
[39] Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, and Wenqi Wei. 2021. Boosting ensemble accuracy by revisiting ensemble diversity metrics. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16469–16477.

[40] Zhongwei Xie, Ling Liu, Lin Li, and Luo Zhong. 2021. Learning joint embedding with modality alignments for cross-modal retrieval of recipes and food images. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2221–2230.

[41] Weilin Xu, David Evans, and Yanjun Qi. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155* (2017).

[42] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. 2021. Towards face encryption by generating adversarial identity masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3897–3907.

[43] Kaiwei Zeng, Munan Ning, Yaohua Wang, and Yang Guo. 2020. Hierarchical clustering with hard-batch triplet loss for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13657–13665.

[44] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.

[45] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[46] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. 2021. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10492–10502.

[47] Thomas Zinner, Oliver Hohlfeld, Osama Abboud, and Tobias Hoßfeld. 2010. Impact of frame rate and resolution on objective QoE metrics. In *2010 second international workshop on quality of multimedia experience (QoMEX)*. IEEE, 29–34.