PRO-Face: A Generic Framework for Privacy-preserving Recognizable Obfuscation of Face Images

Lin Yuan
Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications
Chongqing, China
yuanlin@cqupt.edu.cn

Linguo Liu
Key Lab. of Web Intelligence and Technology, Chongqing University of Posts and Telecommunications
Chongqing, China
linguoliu@foxmail.com

Xiao Pu
Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications
Chongqing, China
puxiao@cqupt.edu.cn

Zhao Li
Key Lab. of Web Intelligence and Technology, Chongqing University of Posts and Telecommunications
Chongqing, China
lizhaot7@126.com

Hongbo Li
Key Lab. of Web Intelligence and Technology, Chongqing University of Posts and Telecommunications
Chongqing, China
lihongbo@cqupt.edu.cn

Xinbo Gao*
Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications
Chongqing, China
gaoxb@cqupt.edu.cn

ABSTRACT

A number of applications rely on automated face recognition to guarantee secure service functioning, and meanwhile, have to take into account the privacy of individuals exposed under camera systems. This is the so-called *Privacy-Utility* trade-off. However, most existing approaches to facial privacy protection focus on removing identifiable visual information from images, leaving protected face unrecognizable to machine, which sacrifice utility for privacy. To tackle the privacy-utility challenge, we propose a novel, generic, effective, yet lightweight framework for **P**rivacy-preserving **R**ecognizable **O**bfuscation of **Face** images (named as **PRO-Face**). The framework allows one to first process a face image using any preferred obfuscation, such as image blur, pixelate and face morphing. It then leverages a Siamese network to fuse the original image with its obfuscated form, generating the final protected image visually similar to the obfuscated one from human perception (for *Privacy*) but still recognized as the original identity by machine (for *Utility*). The framework supports various obfuscations for facial anonymization. The face recognition can be performed accurately not only across anonymized images but also between plain and anonymized ones, based on only pre-trained recognizers. Those feature the "generic" merit of the proposed framework. In-depth objective and subjective evaluations demonstrate the effectiveness of the proposed framework in both privacy protection and utility preservation under distinct scenarios. Our source code, models and supplementary materials are made publicly available[1].

*Corresponding author

[1]https://github.com/fkeufss/PRO-Face

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548202

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Biometrics**; *Computer vision representations.*

KEYWORDS

privacy protection, face obfuscation, face recognition, image fusion

ACM Reference Format:
Lin Yuan, Linguo Liu, Xiao Pu, Zhao Li, Hongbo Li, and Xinbo Gao. 2022. PRO-Face: A Generic Framework for Privacy-preserving Recognizable Obfuscation of Face Images. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548202

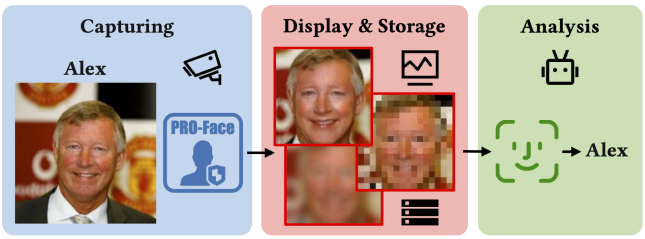

Figure 1: Illustrative example of PRO-Face: Alex's face under camera is immediately protected upon capturing, possibly by different types of obfuscations. Protected images are perceptually anonymized during display and storage, ensuring privacy of Alex. Meanwhile, machine can still identify Alex from protected images, preserving recognition utility.

1 INTRODUCTION

Face recognition has made significant advancements in recent years and is used in a wide range of applications. However, extensive development of such a technology has also raised growing concern on privacy due to the wide spread of individuals faical images. On one hand, certain services require automatic and accurate face recognition to meet security or efficiency needs. On the other hand,

individuals captured by cameras have a common will to protect their facial privacy from intentional or unintentional human observations [45]. Hence, a *Privacy-Utility* balanced approach capable of protecting facial privacy against human eye meanwhile retaining the utility of identity recognition by machine is desirable.

Considerable efforts have been devoted to the research of privacy enhancing technologies (PETs) for face images. Methods studied include conventional privacy filters [1, 4, 25] or transformations [18, 42, 43], adversarial facial perturbations [28, 34, 40] and GAN-based facial anonymizations [2, 11, 16, 19, 20, 24]. While, most existing methods attempt to remove identifiable visual information from images (known as *de-identification* or *anonymization*), leaving protected images unrecognizable to machine. In consequence, there is a lack of solution taking into account both privacy protection and utility preservation.

To tackle the *Privacy-Utility* trade-off, we develop **PRO-Face**, a generic, effective, yet simple framework for **P**rivacy-preserving **R**ecognizable **O**bfuscation of **Face** images. The core of the framework is to fuse key visual information of a face image into its pre-obfuscated version, such that the generated image is visually similar to the obfuscated one (for *Privacy*), but still recognized as its original identity by state-of-the-art recognition models (for *Utility*). The design of PRO-Face is inspired by image steganography and adversarial examples, where image fusion implicitly embeds essential information of original image into to its obfuscated form imperceptibly, and the resulted image attempts to "mislead" a pre-trained face recognizer to make a "wrong" decision towards the original identity although it does not look original. The following "generic" characteristics of the framework make it feasible for a wide range of scenarios:

- The framework supports different types of visual obfuscations, providing high flexibility for users.
- The framework only requires pre-trained face recognizers to fulfill correct recognition on protected images, which could serve as an add-on functionality to an existing facial recognition system.
- The framework allows for identity recognition not only within anonymized faces (anonymized-domain), but also across plain and anonymized faces (cross-domain).

The rest of the paper is structured as follows: Section 2 introduces works related to ours. Section 3 describes in detail the proposed framework, followed by Section 4 presenting the performance evaluations. Finally, Section 5 discusses potential applications and known limitations of the proposed approach, and concludes the paper. The main contributions of this paper are summarized as follows:

- A novel, generic, effective yet lightweight framework capable of generating visually obfuscated but machine recognizable face images is proposed, which could achieve a good balance between privacy and utility.
- An updated Siamese network for image fusion inspired by state-of-the-art deep steganography along with effective training policy based on specialized triplet identity and perceptual loss is studied.
- Extensive objective and subjective evaluations demonstrating both the privacy protection and utility preservation capability of the proposed framework is provided.

2 RELATED WORK

In this section, we review three major categories of research on the topic of *facial privacy protection* that are most relevant to ours.

2.1 Adversarial approaches

Neural networks are known to be vulnerable to adversarial examples [37], which could mislead deep recognition models to generate incorrect results while being imperceptible to human perception. Such a characteristic has been utilized to protect facial attributes from being recognized by unauthorized entities, without affecting the initial usage of original images for viewing. Mirjalili et al. [26–28] proposes a set of methods to conceal soft-biometric attributes (such as gender and race) while preserving the original image quality and recognition utility. Oh et al. [29] introduces a game theory based framework to generate adversarial-perturbed face image that could confound identity recognition of machine. Shan et al. proposes *Fawkes* [34] that alters face with imperceptible perturbations making the facial representation closer to a target face with different identity. Yang et al. [40] proposes a targeted identity-protection iterative method (*TIP-IM*), capable of generating adversarial masks to be overlaid on face images to conceal identity information against face recognizers. Recently, Hu et al. [13] proposes *AMT-GAN* that leverages GAN to synthesize adversarial face images with makeup transferred from reference images. In short, adversarial approaches can well address privacy risks against machine recognition, instead of targeting at human inspection, which contradicts our objective.

2.2 Facial appearance anonymization

Another group of research leverages GAN to edit the facial appearance to achieve anonymization. Sun et al. [35] proposes head inpainting to generate anonymized faces ensuring sensitive information of the original face being removed. Gafni et al. [10] proposes an encoder-decoder architecture capable of anonymizing faces in real-time video by manipulating disentangled feature representations. *DeepPrivacy* [16] utilizes Conditional GAN (CGAN) [31] with background and pose annotation as inputs to guide the generation of realistic anonymized faces. *CIAGAN* [24] leverages an identity-controlling vector and facial key-points to guide a CGAN model to generate face with anonymized visual identity while keeping the original pose. Gu et al. [11] proposes an password-conditioned face identity transformer that performs anonymization and de-anonymization upon a binary-valued password vector. Recently, Cao et al. [2] proposes a personalized and invertible de-identification framework where a user-specific password and an adjustable parameter are used to control the direction and degree of identity variation. However, this category of methods mainly focus on the task of anonymization, without taking much the recognition utility into account.

2.3 Identity-preserving facial anonymization

Different from most existing solutions, we are more interested in anonymizing the facial appearance (to human eye) while preserving the identity recognition capability (by machine), a.k.a. *identity-preserving face anonymization*. Unfortunately, to the best of our knowledge, very few studies have focused on this specific problem.

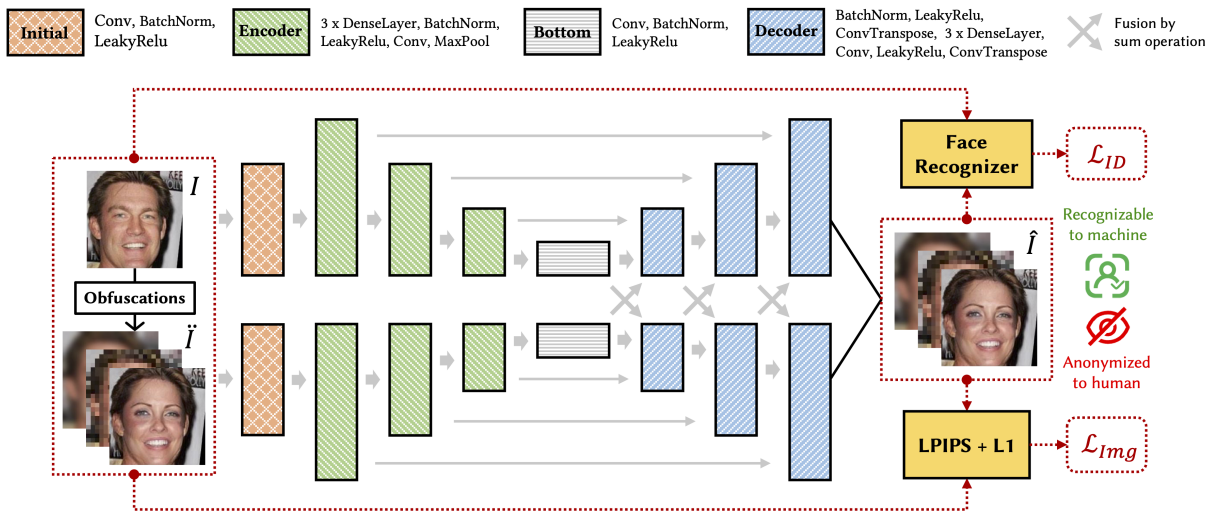

Figure 2: Illustration of PRO-Face framework: Input image I is processed by a selected obfuscation (blur, pixelate or face morphing), resulting in pre-obfuscated image $\ddot{I}$. I and $\ddot{I}$ are sent to a Siamese network and fused in different scales of feature maps in decoder stage, generating PRO-Face protected image $\hat{I}$. The training takes I and $\ddot{I}$ as reference to supervise the generation of $\hat{I}$, with respect to a pre-trained face recognizer. Specialized triplet identity loss $\mathcal{L}_{ID}$ and image loss $\mathcal{L}_{Img}$ are utilized to optimize the fusion image towards the pre-obfuscated image perceptually, while staying close to the original image in embedding space.

Li et al. [21] proposes an identity-preserving face camouflage approach, which utilizes an encoder-decoder network to disentangle facial features into an appearance code and an identification code, where the former is replaced according to another target image. Recently, the same author [20] proposes a more advanced solution to the same problem via adaptive facial attributes obfuscation. The method first relies on an identity-aware region discovery module to determine the facial attributes sensitive to human eye, and then carries out anonymization via a conditional GAN model taking the original face and attribute indicator as input. However, none study achieves practicable recognition accuracy.

3 THE FRAMEWORK

Our framework operates in an intuitive and straightforward manner: Given an input face image I, it first allows one to pre-obfuscate it using any preferred operation, such as blur, pixelate and face morphing. The pre-obfuscated image is noted as $\ddot{I}$. Then, both the original I and the pre-obfuscated $\ddot{I}$ are fed into a deep fusion network, generating the final obfuscated face $\hat{I}$, which highly looks like the pre-obfuscated $\ddot{I}$. Meanwhile, a pre-trained face recognizer can still extract the similar identity representation from $\ddot{I}$ as the original I, so as to perform face recognition accurately on the obfuscated image $\ddot{I}$. In such a manner, the framework protects the facial privacy of the individual in image against human visual perception and meanwhile preserves the utility for automatic face recognition.

3.1 Intuition

Before diving into technical details, we would like to provide our intuitions and inspirations for designing such a framework:

Simulating protected image as adversarial example. We are first inspired by *Fawkes* [34], which creates adversarial images to "fool" existing face recognizers. The adversarial image is close to a target image (with different identity) in embedding space but similar to the original image in terms of human visual perception. This inspires us to design an opposite mechanism with switched inputs, such that the generated image looks like a visually obfuscated image as target while being extracted the original face embedding "incorrectly" by an existing recognizer. However, our objective is more challenging as we expect the face recognizer to "misbehave" consistently towards a specific identity, while *Fawkes* [34] only expects the recognizer to make mistakes towards any other identity.

Simulating protection as image steganography. Another inspiration for us is from the research of image steganography, such as Light Field Messaging (LFM) [39], where a message image is embedded into a carrier image imperceptibly via a deep neural network, and the former could be later detected from the encoded image by another network. It motivates us to embed critical information of a face image into its obfuscated form imperceptibly such that the identity information is still detectable. Differently, the "detector" in our case is simply a pre-trained face recognizer, which is never updated specifically for our task. This is more challenging but makes our approach more usable.

3.2 Network design

Our network is designed by adopting the steganography architecture proposed in [39], which follows a Siamese network [17] structure, composed of two subnets receiving the original image and its pre-obfuscated version respectively. The two subnets share the same structure but with different weights. Each subnet of the Siamese network is featured with a U-Net [32] architecture, where both the encoder and decoder consists of multiple Dense blocks [14], featuring image maps at different scales. We make the image fusion

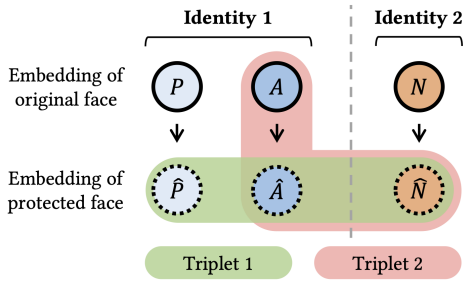

Figure 3: Specialized triplet construction for optimizing face matching in anonymous image domain (Triplet 1) and original/anonymous cross-domain (Triplet 2) respectively.

happen during the decoding stage of the U-Net, where the outputs of each block (including the bottom layer) from the two subnets are summed up before being sent to subsequent blocks. This is a key difference compared with [39], which merges the feature maps at both encoding and decoding stage. The intuition for this change is to take less visual information from the original image as possible during fusion. The details of the network components along with the training strategy are illustrated in Figure 2.

3.3 Training strategy and loss functions

We train the fusion network each time with a specific type of obfuscation with respect to a pre-trained face recognition model. The training is done by optimizing the network parameters so that the output image is perceptually similar to pre-obfuscated one while its face embedding extracted by the recognition model is close to the original image. Therefore, the training employs two types of losses: 1) the *Identity loss* for optimizing the face embedding and 2) the *Image loss* for optimizing the facial appearance. To improve model generalization, we vary obfuscation strength/variants over training batches, such as changing blur kernel, pixelate block size, and target face for morphing. Details of the losses are given below.

Identity loss. We first sample a set of triplet images, consisting of a pair of positive/anchor sample with the same identity, and a negative sample with a different identity, noted as I_P, I_A and I_N respectively. Having their corresponding PRO-Face protected images, noted as $\hat{I}_P$, $\hat{I}_A$ and $\hat{I}_N$, our identity loss is defined as the sum of two triplet losses:

$$\mathcal{L}_{ID} = \mathcal{L}_T(\hat{I}_A, \hat{I}_P, \hat{I}_N) + \mathcal{L}_T(\hat{I}_A, I_A, \hat{I}_N), \tag{1}$$

where $\mathcal{L}_T$ is a standard triplet loss applied on the embedding space $E(\cdot)$ of triplet samples with $L2$ as distance metric and α as margin:

$$\mathcal{L}_T(A, P, N) = \|E(A) - E(P)\|_2 - \|E(A) - E(N)\|_2 + \alpha \tag{2}$$

Figure 3 illustrates the two losses defined in Equation (1). The former tries to separate embeddings of anonymized triplet samples, in order to improve the discrimination performance in anonymized image domain. The later tries to optimize the embedding of anonymized image towards its original version, instead of the negative sample, which aims to improve the face matching performance between original and anonymized images (cross-domain).

Face recognizer	Train. method	# param.	Acc. LFW
MobileFaceNet [6]	ArcFace [8]	1M	0.9863
InceptionResNet [36]	FaceNet [33]	28M	0.9906
IResNet50 [9]	ArcFace [8]	44M	0.9898
SEResNet50 [12]	ArcFace [8]	44M	0.9896
IResNet100 [9]	ArcFace [8]	65M	0.9983

Table 1: Details of pre-trained face recognizers in evaluation (named by their backbone), including the training method, number of model parameters and the accuracy on LFW.

Image loss. To generate fusion image visually obfuscated, we define the visual triplet loss as follows:

$$\mathcal{L}_{Visual} = \mathcal{L}_{Perc}(\ddot{I}, \hat{I}) - \mathcal{L}_{Perc}(I, \hat{I}) + \beta, \tag{3}$$

where $\mathcal{L}_{Perc}$ measures perceptual similarity between two images, such as the LPIPS [44]. In addition, we also use the $L1$ distance to constrain image similarity in pixel level:

$$\mathcal{L}_{L1} = \|\ddot{I} - \hat{I}\|_1. \tag{4}$$

Finally, the overall objective function for training our model is defined as the weighted sum of the above losses:

$$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_{ID} + \lambda_2 \mathcal{L}_{Visual} + \lambda_3 \mathcal{L}_{L1} \tag{5}$$

4 EXPERIMENTS

We conduct extensive experiments to verify the effectiveness of the proposed approach in privacy protection and utility preservation. To evaluate privacy, we objectively compute similarity metric to check the visual difference between protected image with its original, and subjectively conduct user study to inspect how the proposed obfuscation could confound human's recognition capability. To evaluate utility, we comprehensively analyze face verification performance under distinct scenarios.

4.1 Experimental setup

4.1.1 Datasets. We use the following datasets in experiments:

CelebFaces Attributes Dataset (CelebA) [23] is one of the most widely used datasets for face recognition, containing 202'599 celebrity images of 10'177 identities. Each image is annotated with 5 landmarks and 40 binary attributes. The dataset is split in training, validation and testing sets.

VGGFace2 [3] contains more than 3 million images from 9'131 identities, obtained mainly from Google Image Search. The images are with larger variations in pose, age, illumination, ethnicity and profession. This dataset is divided into the training and test set, each containing 8'631 and 500 identities respectively.

Labeled Faces in the Wild (LFW) [15] contains more than 13'000 images of 5'749 identities and provides a standard benchmark for face verification, composed of 6'000 face pairs with 3'000 matched and the other 3'000 non-matched.

We crop and align all images to keep the head region in center with resolution of 112×112. The cropping and alignment are implemented either with pre-annotated facial landmarks (for CelebA) or by modern models[2] [7]. We use training images of CelebA to

[2] https://insightface.ai/

Obfuscation ▷	Blur		Pixelate		FaceShifter [22]		SimSwap [5]	
Recognizer ▽	LPIPS ↑	SSIM ↓	LPIPS ↑	SSIM ↓	LPIPS ↑	SSIM ↓	LPIPS ↑	SSIM ↓
MobileFaceNet [6]	0.444	0.624	0.637	0.546	0.121	0.836	0.162	0.854
InceptionResNet [36]	0.430	0.660	0.633	0.597	0.111	0.854	0.149	0.872
IResNet50 [9]	0.435	0.644	0.637	0.527	0.115	0.854	0.157	0.862
SEResNet50 [12]	0.450	0.626	0.638	0.551	0.115	0.852	0.158	0.853
IResNet100 [9]	0.428	0.638	0.633	0.593	0.112	0.845	0.150	0.875

Table 2: Objective measurement of privacy by similarity between protected and original images (on LFW).

Method ▽ | Dataset ▷	CelebA	VGGFace2
Li (Adaptive) [20]	0.114	0.134
PRO-Face (Blur)	0.464 (+0.350)	0.447 (+0.313)
PRO-Face (Pixelate)	0.637 (+0.523)	0.636 (+0.502)
PRO-Face (FaceShifter)	0.103 (−0.011)	0.120 (−0.014)
PRO-Face (SimSwap)	0.170 (+0.054)	0.179 (+0.045)

Table 3: Comparison with Li's Adaptive method [20] on CelebA and VGGFace2 in terms of LPIPS. For our method, the average LPIPS over the five recognizers is used.

Image type	Accuracy	Confidence
Original	0.920	4.20
PRO-Face (Blur)	0.675 (−0.245)	3.55 (−0.65)
PRO-Face (Pixelate)	0.520 (−0.400)	3.30 (−0.9)
PRO-Face (FaceShifter)	0.675 (−0.245)	3.67 (−0.53)
PRO-Face (SimSwap)	0.700 (−0.220)	3.79 (−0.41)

Table 4: Subjective recognition accuracy and average confidence score for original and protected images. The confidence score is rated as integer ranging from 1 to 5, meaning *Unsure, Not very sure, Neutral, Sure* and *Very sure* respectively.

train our models, which are tested on every dataset: For LFW, we use the provided benchmark as it is; For CelebA and VGGFace2, we each construct a benchmark set by randomly sampling 6'000 image pairs (similar as LFW) from their test split.

4.1.2 Face recognizers. We experiment with five face recognition models with different backbones, from the simple MobileFaceNet [6] to more sophisticated IResNet100 [9]. More details are shown in Table 1. All models generate face embedding in same dimension of 512. They were trained on a different image dataset (CISIA-Webface [41]) and achieve satisfactory recognition accuracy on all above datasets.

4.1.3 Obfuscations. We experiment with four image obfuscations: blur, pixelate and two types of face morphing. For blur, we use Gaussian with kernel size of 31 and standard deviation (σ) from 2 to 8 in training (fixed $\sigma = 5$ in testing). For pixelate, we vary block size from 4 to 10 in training and fixed it as 7 in testing. For face morphing, we use two recent face swapping algorithms, FaceShifter [22] and SimSwap [5], to morph the facial appearance of an image towards a target face. We split the validation set of CelebA in two halves to offer target faces for training and testing phases respectively.

4.1.4 Implementation details. We train a model w.r.t. every combination of obfuscation and pre-trained face recognizer, resulting in a total of 20 models (4 obfuscations × 5 recognizers). The face recognition models are not updated during training. For blur and pixelate, we vary the parameter σ and block size randomly over batches. For face morphing, we randomly sample a single target image for each batch. The training is done using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate of 0.001. Each training batch contains 8 randomly sampled triplets, resulting in 24 images per batch.

4.2 Evaluation on privacy protection

4.2.1 ***Objective evaluation.*** Objectively, we utilize LPIPS [44] and SSIM [38] to measure the similarity between the protected and original image. Higher LPIPS (or lower SSIM) indicates stronger privacy protection capability. The results for different combinations of obfuscation-recognizer for LFW dataset are shown in Table 2. We observe that all models result in relatively high privacy score, where blur and pixelate exhibit stronger protection than the other two, which is reasonable as blur and pixelate are applied on the entire image and change pixels drastically.

Specifically, we compare our method with Li's Adaptive approach [20] on CelebA and VGGFace2 datasets in terms of LPIPS (Li [20] provides no results on LFW). Since we don't have access to the implementation of Li's method and are unaware of the exact recognition model applied, we only compare the average LPIPS over different recognizers in our case with [20]. As is shown in Table 3, our approach with blur and pixelate outperforms [20] by inducing LPIPS gain from 0.313 to 0.523. For face morphing, our LPIPS scores are comparable to [20].

4.2.2 ***Subjective evaluation.*** Subjectively, we first show in Figure 4 several example images protected with different obfuscations w.r.t. IResNet100. Qualitatively, we observe that: 1) For each obfuscation, the protected image by fusion is visually somewhere between the original and pre-obfuscated image, closer to the pre-obfuscated one, thus enhancing visual privacy; 2) the degree of visual anonymization highly depends on the obfuscation applied, where stronger obfuscation usually results in stronger anonymization, such as pixelate.

To further verify the privacy protection capability, we conduct a user study via online crowdsourcing to evaluate how the proposed obfuscations could confound human's recognition ability. Different from Li [20], we put subjects in a more realistic scenario by asking them to recognize face images in different forms given a set of reference images. We randomly select 40 identities (20 males and 20 females) from LFW. For each identity, we randomly sample a test image and leave the remaining as reference set. Five types of each test image are evaluated, including their original form and four

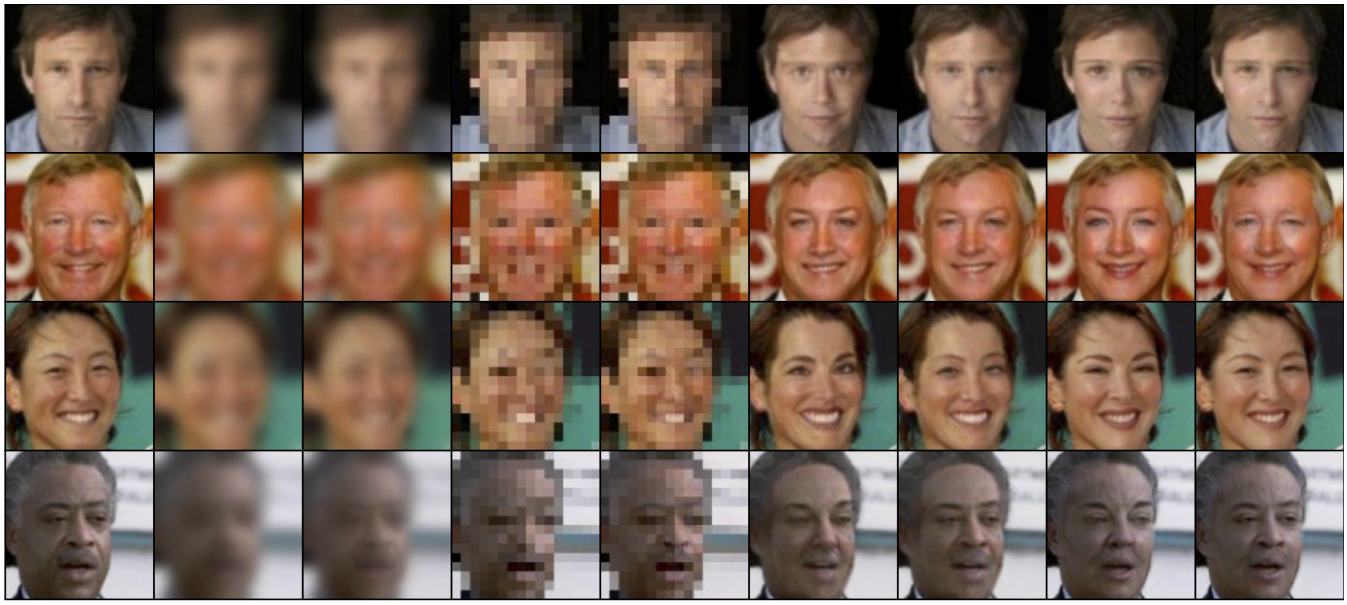

Figure 4: Example images in original and different obfuscated forms. The symbol ⋆ means PRO-Face protected w.r.t. IResNet100.

Obfuscation ▷	Blur		Pixelate		FaceShifter		SimSwap	
Recognizer ▽	ADR	XDR	ADR	XDR	ADR	XDR	ADR	XDR
MobileFaceNet	0.833/0.943	0.859/0.950	0.873/0.951	0.874/0.954	0.911/0.973	0.759/0.915	0.914/0.972	0.781/0.921
InceptionResNet	0.888/0.962	0.929/0.968	0.831/0.948	0.886/0.960	0.940/0.974	0.934/0.969	0.952/0.975	0.942/0.971
IResNet50	0.867/0.956	0.894/0.953	0.874/0.947	0.883/0.954	0.941/0.977	0.889/0.951	0.933/0.975	0.808/0.936
SEResNet50	0.821/0.948	0.859/0.953	0.815/0.934	0.811/0.937	0.927/0.977	0.890/0.951	0.924/0.972	0.824/0.933
IResNet100	0.945/0.971	0.959/0.966	0.890/0.965	0.940/0.964	0.974/0.988	0.959/0.969	0.971/0.991	0.952/0.967

Table 5: Face verification performance of PRO-Face in terms of TAR w.r.t. different obfuscations and recognizers on LFW. Two values split by / indicate TAR @ FAR = 0.01 and 0.1 respectively. ADR/XDR indicate Anonymized-/Cross-domain respectively.

PRO-Face protected forms (model w.r.t. IResNet100). We conduct experiments on Amazon Mechanical Turk (MTurk)[3], where each Human Intelligence Task (HIT) presents a probe image (sampled from test set) along with nine candidate images (from reference set). The nine candidates include an image with the same identity as probe and eight images from other identities. Subjects are required to select from candidate images the one that best matches the probe in terms of identity, along with their confidence. We require each HIT to be rated by at least 5 subjects, and randomly inject "honeypot" HITs where the probe repeats one of the candidates for filtering sloppy subjects. We reward each subject 0.02 US Dollars for completing every HIT successfully. Finally we collect answers from 1'180 HITs rated by 37 workers, each completing 32 HITs in average. For each image type, we compute the overall recognition accuracy (proportion of correctly recognized HITs) and corresponding mean confidence score, shown in Table 4. The results indicate that images protected with PRO-Face all give reduced recognition performance in both accuracy and confidence level compared with original. Nevertheless, the degree of performance reduction varies depending on obfuscation applied: pixelate again results in the strongest protection effect, in line with our objective evaluations in Section 4.2.1.

[3]https://www.mturk.com/

4.3 Evaluation on utility preservation

To evaluate the performance of utility preservation, we run face verification benchmark on different trained models. Here we consider two different recognition scenarios:

- **Anonymized-Domain Recognition (ADR):** All images under recognition are protected by PRO-Face. This scenario aims to evaluate face matching within anonymized images.
- **Cross-Domain Recognition (XDR):** Each pair of images under verification consists of one in original and the other in protected form. This scenario aims to evaluate face matching between plain and anonymized images.

The verification results in terms of True Accept Rate (TAR) at two False Accept Rate (FAR) values (0.01 and 0.1) for LFW dataset is shown in Table 5. According to the results, all models result in satisfactory verification rates in both ADR and XDR scenarios.

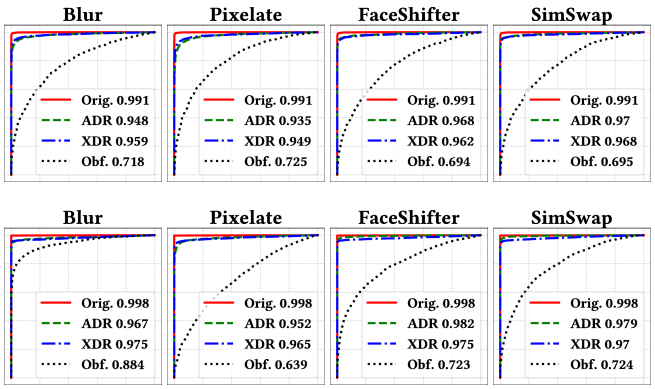

Figure 5: ROC curves and Accuracy (in legend) of InceptionResNet (1st row) and IResNet100 (2nd row) on LFW under different image domains: original (Orig.), anonymized-domain (ADR), cross-domain (XDR) and pre-obfuscated (Obf.).

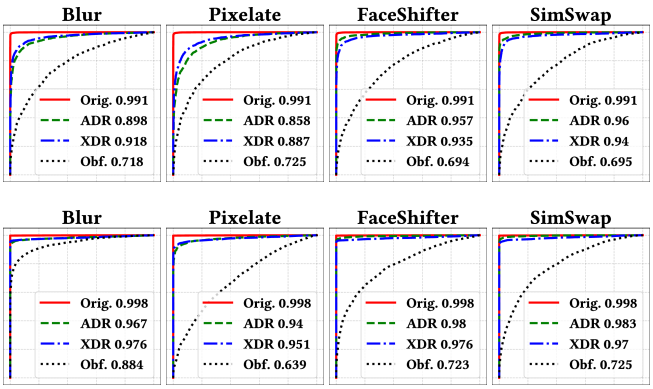

Figure 6: Illustration of transferability effect by recognition ROC and Accuracy (on LFW) in different scenarios. The 1st row: performance of InceptionResNet on protection model trained with IResNet100; The 2nd row: the opposite.

(a) CelebA

	Li (Adaptive) [20]		PRO-Face (SimSwap)	
	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1
Min.	0.644	0.752	**0.868**	**0.947**
Max.	0.784	0.856	**0.937**	**0.978**
Avg.	0.720	0.816	**0.898**	**0.961**

(b) VGGFace2

	Li (Adaptive) [20]		PRO-Face (SimSwap)	
	FAR=0.01	FAR=0.1	FAR=0.01	FAR=0.1
Min.	0.479	0.731	**0.644**	**0.842**
Max.	0.613	0.771	**0.783**	**0.939**
Avg.	0.537	0.749	**0.690**	**0.884**

Table 6: Comparison between our method (SimSwap) with Li's Adaptive method [20] in terms of Min./Max./Avg. TAR@FAR=0.01/0.1 over different face recognizers in ADR scenario on (a) CelebA and (b) VGGFace2.

To prove the superiority of our approach, we again compare with Li [20] on CelebA and VGGFace2. For the same reason mentioned earlier, we only compare the minimum, maximum and average TAR over different recognizers between two methods. In our case, ADR scenario with obfuscation SimSwap is used for comparison, as Li [20] only supports verification within anonymized images and its visual protection effect is similar to SimSwap. Table 6 shows the comparison results for CelebA and VGGFace2 respectively, from which our min. TAR is even higher than the max. TAR of [20]. Considering PRO-Face (SimSwap) exhibits comparable privacy score as [20] (see Table 3), the advantage of our approach is significant.

Effectiveness of image fusion. One may argue that state-of-the-art face recognizers may already recognize pre-obfuscated images with a good level of accuracy, so it is not clear how much our fusion mechanism helps with recognition. To verify the effectiveness of the proposed fusion mechanism, we also inspect the verification accuracy on obfuscated images without fusion applied. We compute the Receiver Operator Characteristic (ROC) curve and Accuracy of verification for original, anonymized-domain, cross-domain and pre-obfuscated images respectively. The results for two selected face recognizers (InceptionResNet and IResNet100) are shown in Figure 5. One observes that the recognition performance on pre-obfuscated images without fusion is relatively poor (close to random guess for certain cases). With PRO-Face fusion applied, the recognition rate improves greatly towards the original cases.

Transferability. Above evaluations assume that the test stage uses the same face recognizer as is used to train the model. Under more general scenarios, the effectiveness of our framework is further enhanced relying on the *transferability* effect, the property that models trained for similar tasks share similar properties, even when they were trained on different architectures [30]. This suggests that the recognition should still be effective even if the fusion network is trained w.r.t. a different recognition model. To verify the transferability effect, we evaluate the verification performance of IResNet100 on model trained with InceptionResNet, and vice versa. The verification results represented by ROC and accuracy on LFW for both scenarios are show in Figure 6. In both model transferring scenarios (IResNet100⇌InceptionResNet), the recognition performance remains high (accuracy greater than 0.9 for most cases), which well demonstrates the recognition transferability of our framework.

Privacy-utility trade-off. Last but not least, we embrace a short evaluation on the trade-off between privacy and utility of the proposed approach. To observe such trade-off, we vary the pixelate strength in test stage and evaluate the recognition performance versus the image similarity metric. More specifically, we use different pixelate block sizes ranging from 4 to 12 to generate PRO-Face images using the fusion model trained w.r.t IResNet100. Then we compute the image similarity by LPIPS (PRO-Face vs original) and the verification accuracy of IResNet100 on LFW. The scatter plot

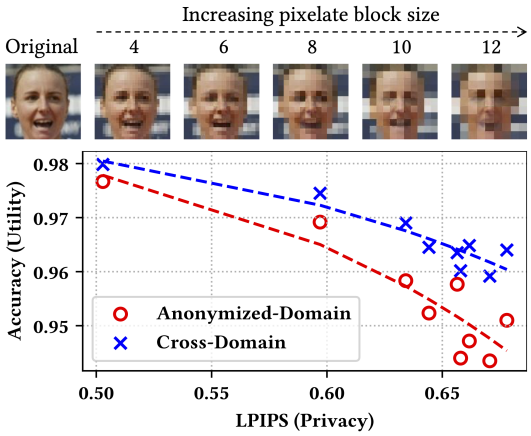

Figure 7: Illustration of Privacy-Utility trade-off featured by LPIPS vs recognition accuracy.

of anonymized-/cross-domain recognition accuracy versus LPIPS is shown in Figure 7. One observes a clear decrease in recognition accuracy (utility) along with the increase of LPIPS (privacy). Nevertheless, even with the strongest pixelate applied, in which case the facial appearance is highly obscured, the recognition performance still remains effectively high (>0.94 in accuracy).

Lastly, pre-trained models (ours and face recognizers), full results on CelebA and VGGFace2, more resulted images and exemplary MTurk HITs can be found at our GitHub page[4].

5 DISCUSSION AND CONCLUSION

In this paper, we present a generic framework for protecting visual privacy of face images while preserving a good level of utility for automatic identity recognition. The framework follows a concept of **P**rivacy-preserving **R**ecognizable **O**bfuscation for **Face** images, and is therefore named **PRO-Face**. The core of the PRO-Face framework is to fuse critical identifiable information of a face image with its pre-obfuscated version via a Siamese network, such that the resulted fusion image is perceptually similar to the pre-obfuscated one (ensuring privacy), but can still be recognized as the original identity by state-of-the-art face recognizers (retaining utility). To train the fusion network, we design specialized triplet loss to meet the privacy and utility requirements with model optimization. Objective and subjective evaluations demonstrate the effectiveness of the proposed approach in enhancing visual privacy while preserving a high level of recognition performance.

The framework is featured with the following merits: First, the framework supports a number of obfuscation effects to ensure visual privacy, including but not limited to image blur, pixelate and face morphing, verified in our experiments. This would provide potential users with flexibility of choosing their preferred obfuscation in application. Second, the preserved recognition utility under this framework only relies an existing pre-trained face recognizer and shows some level of transferability. This makes it possible to use the framework as a simple add-on functionality to an existing facial recognition system so as to minimize the burden for system engineering. Third, accurate face recognition can be achieved not only in anonymized-domain (face matching between protected images) but also in cross-domain (between plain and protected images). This allows for privacy-preserving face recognition in broader scenarios. Last but not least, the model only has 0.2M parameters (5× smaller than the minimum face recognizer MobileFaceNet), which would not create much overhead to real-time performance in application.

Regarding the potential applications of PRO-Face, we would like to describe two scenarios: The first is video surveillance, where the facial parts of all individuals captured by surveillance camera could be protected by the proposed approach, such that unauthorized people would only see anonymized faces. However, the surveillance system can still recognize each person in a real-time manner to perform critical tasks such as statistics, security alert or forensics. Another scenario is face authentication system, where all faces being captured, displayed and stored can all be kept in protected form. Meanwhile, the protected faces can still accomplish the authentication task correctly due to the preserved recognition utility. Additionally, the authentication can be done regardless of the form of the template face, as the recognition under our framework is applicable in both anonymized- and cross-domain scenarios. This is particularly useful even if the template faces have to be in unprotected plain form by local policy.

Frankly speaking, several limitations of the proposed approach still remain. For instance, the protected image with fusion mechanism visually sits between the original and the pre-obfuscated form, which still reveals certain amount of original visual features that may threat privacy. The privacy protection capability highly relies on the obfuscation applied, e.g. pixelate demonstrates the strongest protection while the protection capability of face morphing depends on the target image selected. This calls for more elaborate mechanisms for selecting satisfactory obfuscations. The recognition robustness of the protected image against commonly applied image degradation (filtering, compression, etc.) is not addressed in this paper, although it is important as it highly impacts the utility of the proposed method. Moreover, same as most other approaches to facial anonymization, only obfuscating the facial region may not guarantee rigid privacy protection, since non-facial features (e.g., skin color, hair style and clothing) could still provide visual cues to disclose privacy. This issue has also been found during our subjective experiments. Therefore, solutions to solve the above limitations will serve as our future work.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 62036007, and in part by the Special Project on Technological Innovation and Application Development under Grant No. cstc2020jscx-dxwtB0032, the Chongqing Excellent Scientist Project under Grant No. cstc2021ycjh-bgzxm0339, the Opening Project of GuangDong Province Key Laboratory of Information Security Technology under Grant No. 2020B1212060078, the Opening Project of Intelligent Policing Key Laboratory of Sichuan Province under Grant No. ZNJW2022KFQN006, and the Science and Technology Research Program of Chongqing Municipal Education Commission under Grant No. KJQN201900628.

[4]https://github.com/fkeufss/PRO-Face

REFERENCES

[1] Prachi Agrawal and P. J. Narayanan. 2011. Person De-Identification in Videos. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 21, 3 (2011), 299–310.

[2] Jingyi Cao, Bo Liu, Yunqian Wen, Rong Xie, and Li Song. 2021. Personalized and Invertible Face De-Identification by Disentangled Identity Information Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2021)*. 3334–3342.

[3] 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)*. 67–74.

[4] Ankur Chattopadhyay and T.E. Boult. 2007. PrivacyCam: A Privacy Preserving Camera Using uCLinux on the Blackfin DSP. In *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*. 1–8.

[5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of the 28th ACM International Conference on Multimedia (MM 2020)*. 2003–2011.

[6] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In *Chinese Conference on Biometric Recognition (CCBR 2018)*. 428–438.

[7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*. 5202–5211.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. 4685–4694.

[9] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. 2021. Improved Residual Networks for Image and Video Recognition. In *2020 25th International Conference on Pattern Recognition (ICPR 2020)*. 9415–9422.

[10] Oran Gafni, Lior Wolf, and Yaniv Taigman. 2019. Live Face De-Identification in Video. In *IEEE/CVF International Conference on Computer Vision (ICCV 2019)*, Vol. 00. 9377–9386.

[11] Xiuye Gu, Weixin Luo, Michael S. Ryoo, and Yong Jae Lee. 2020. Password-Conditioned Anonymization and Deanonymization with Face Identity Transformers. In *European Conference on Computer Vision (ECCV 2020)*. 727–743.

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 42, 8 (2020), 2011–2023.

[13] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)*.

[14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*. 2261–2269.

[15] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

[16] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. DeepPrivacy: A Generative Adversarial Network for Face Anonymization. In *International Symposium on Visual Computing (ISVC 2019)*. 565–578.

[17] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese Neural Networks for One-shot Image Recognition. In *ICML Deep Learning Workshop*.

[18] Pavel Korshunov and Touradj Ebrahimi. 2013. Using Warping for Privacy Protection in Video Surveillance. In *2013 18th International Conference on Digital Signal Processing (DSP)*. 1–6.

[19] Zhenzhong Kuang, Huigui Liu, Jun Yu, Aikui Tian, Lei Wang, Jianping Fan, and Noboru Babaguchi. 2021. Effective De-identification Generative Adversarial Network for Face Anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3182–3191.

[20] Jingzhi Li, Lutong Han, Ruoyu Chen, Hua Zhang, Bing Han, Lili Wang, and Xiaochun Cao. 2021. Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM 2021)*. 3891–3899.

[21] Jingzhi Li, Lutong Han, Hua Zhang, Xiaoguang Han, Jingguo Ge, and Xiaochun Cao. 2021. Learning Disentangled Representations for Identity Preserving Surveillance Face Camouflage. In *2020 25th International Conference on Pattern Recognition (ICPR 2020)*. 9748–9755.

[22] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing High Fidelity Identity Swapping for Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*. 5073–5082.

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV 2015)*. 3730–3738.

[24] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. 2020. CIAGAN: Conditional Identity Anonymization Generative Adversarial Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020)*, Vol. 00. 5446–5455.

[25] Blaž Meden, Peter Rot, Philipp Terhörst, Naser Damer, Arjan Kuijper, Walter J. Scheirer, Arun Ross, Peter Peer, and Vitomir Štruc. 2021. Privacy–Enhancing Face Biometrics: A Comprehensive Survey. *IEEE Transactions on Information Forensics and Security (TIFS)* 16 (2021), 4147–4183.

[26] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2018. Gender Privacy: An Ensemble of Semi Adversarial Networks for Confounding Arbitrary Gender Classifiers. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS 2018)*, Vol. 00. 1–10.

[27] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2019. FlowSAN: Privacy-Enhancing Semi-Adversarial Networks to Confound Arbitrary Face-Based Gender Classifiers. *IEEE Access* 7 (2019), 99735–99745.

[28] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. 2020. PrivacyNet: Semi-Adversarial Networks for Multi-Attribute Face Privacy. *IEEE Transactions on Image Processing (TIP)* 29 (2020), 9400–9412.

[29] Seong Joon Oh, Mario Fritz, and Bernt Schiele. 2017. Adversarial Image Perturbation for Privacy Protection A Game Theory Perspective. In *2017 IEEE International Conference on Computer Vision (ICCV 2017)*. 1491–1500.

[30] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. https://arxiv.org/abs/1605.07277

[31] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR 2016)*.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*. 234–241.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. 815–823.

[34] 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.

[35] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Natural and Effective Obfuscation by Head Inpainting. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. 5050–5059.

[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2018)*. 4278–4284.

[37] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing Properties of Neural Networks. In *2014 International Conference on Learning Representations (ICLR 2014)*.

[38] Z. Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale Structural Similarity for Image Quality Assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, Vol. 2. 1398–1402 Vol.2.

[39] Eric Wengrowski and Kristin Dana. 2019. Light Field Messaging With Deep Photographic Steganography. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019)*. 1515–1524.

[40] 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 *IEEE/CVF International Conference on Computer Vision (ICCV 2021)*. 3897–3907.

[41] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Learning Face Representation from Scratch. https://arxiv.org/abs/1411.7923

[42] Lin Yuan and Touradj Ebrahimi. 2017. Image Privacy Protection with Secure JPEG Transmorphing. *IET Signal Processing* 11, 9 (2017), 1031–1038.

[43] Lin Yuan, Pavel Korshunov, and Touradj Ebrahimi. 2015. Secure JPEG Scrambling Enabling Privacy in Photo Sharing. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 04. 1–6.

[44] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018)*. 586–595.

[45] Shikun Zhang, Yuanyuan Feng, and Norman Sadeh. 2021. Facial Recognition: Understanding Privacy Concerns and Attitudes Across Increasingly Diverse Deployment Scenarios. In *Seventeenth Symposium on Usable Privacy and Security (USENIX SOUPS 2021)*. USENIX Association, 243–262.