Adversarial Privacy-preserving Filter

Jiaming Zhang[1,2], Jitao Sang[1,2], Xian Zhao[1], Xiaowen Huang[1], Yanfeng Sun[3], Yongli Hu[3]
[1]School of Computer and Information Technology & Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[2]Peng Cheng Laboratory, ShenZhen, China
[3]Beijing Key Laboratory of Multimedia and Intelligent Software Technology & Beijing Artificial Intelligence Institute, Faculty of Information Technology, Beijing University of Technology, Beijing, China
lanzhang1107@gmail.com,{jtsang,20120454,xwhuang}@bjtu.edu.cn,{yfsun,huyongli}@bjut.edu.cn

ABSTRACT

While widely adopted in practical applications, face recognition has been critically discussed regarding the malicious use of face images and the potential privacy problems, e.g., deceiving payment system and causing personal sabotage. Online photo sharing services unintentionally act as the main repository for malicious crawler and face recognition applications. This work aims to develop a privacy-preserving solution, called Adversarial Privacy-preserving Filter (APF), to protect the online shared face images from being maliciously used. We propose an end-cloud collaborated adversarial attack solution to satisfy requirements of privacy, utility and non-accessibility. Specifically, the solutions consist of three modules: (1) image-specific gradient generation, to extract image-specific gradient in the user end with a compressed probe model; (2) adversarial gradient transfer, to fine-tune the image-specific gradient in the server cloud; and (3) universal adversarial perturbation enhancement, to append image-independent perturbation to derive the final adversarial noise. Extensive experiments on three datasets validate the effectiveness and efficiency of the proposed solution. A prototype application is also released for further evaluation. We hope the end-cloud collaborated attack framework could shed light on addressing the issue of online multimedia sharing privacy-preserving issues from user side.[1]

CCS CONCEPTS

• **Security and privacy → Privacy protections**.

KEYWORDS

privacy-preserving, face recognition, adversarial example, photo sharing

ACM Reference Format:
Jiaming Zhang[1,2], Jitao Sang[1,2], Xian Zhao[1], Xiaowen Huang[1], Yanfeng Sun[3], Yongli Hu[3]. 2020. Adversarial Privacy-preserving Filter. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413906

[1] To encourage reproducible research, the code is available at GitHub. Furthermore, a demo video illustrates the proposed APF solution in YouTube.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413906

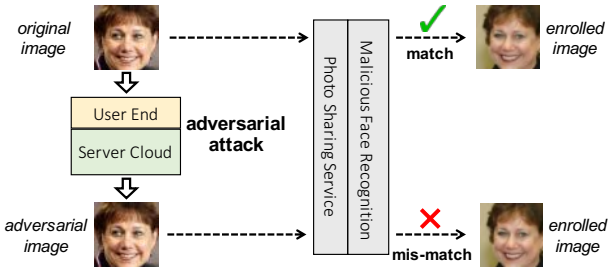

Figure 1: Schematic illustration of the proposed adversarial privacy-preserving filter. Given a face image, the synthetic adversarial image is expected to fool the malicious face recognition algorithm.

1 INTRODUCTION

Benefited from both algorithmic development and the adequate face image data, face recognition has been widely adopted in applications like criminal monitoring, security unlock, digital ticket and even face payment. However, the discussion of privacy problems caused by the unreasonable use of face images has never been stopped. Kneron tested that widely used face payment systems like AliPay and WeChat can be fooled by masks and face images, and Deepfake creating imitated videos causes severe personal sabotage [2]. Among the sources of personal face images, online photo sharing services unintentionally act as the main repository for malicious crawler and face recognition applications. Recently, a company called Clearview was disclosed to crawl face images from Facebook, YouTube and other websites, and construct a database containing more than three billion images. Therefore, it is critical to design privacy-preserving solutions for photo sharing services to protect the uploaded face images from being maliciously used.

This paper aims to preserve users' portrait privacy without affecting their photo sharing experience. Two requirements need to be first satisfied: (1) *privacy*, that user's identification information is unrecognizable from the shared face images; (2) *utility*, that the face image quality should not be undermined. It is interesting to note that adversarial attack [31] exactly meets the above requirements: (1) the direct consequence of adversarial attack is to mislead model predictions which is consistent to fool the malicious face recognition algorithms; and (2) the introduced adversarial perturbation

is trivial and imperceptible to human, thus retaining the image perception and utility for photo sharing. Therefore, we exploit adversarial attack to design the privacy-preserving filter, which adds adversarial perturbation to the original image before uploading to photo sharing services and expects to mislead the malicious face recognition algorithms (illustrated in Fig. 1).

While sharing the adversarial face image guarantees privacy-preserving on photo sharing services, there still remains risk of privacy leakage when generating the adversarial images. Since most adversarial attack solutions involve with complex models and massive computation operations, it is typical to upload the original face images to the privacy-preserving filter server for processing. This inevitably exposes the original face image to the third-party application and makes them vulnerable both during transmission and in the cloud. Therefore, a third requirement is needed for privacy-preserving filter: *non-accessibility*, that the original face image should only be accessible in the user end. To address this, we propose an end-cloud collaborated adversarial attack solution, where the original face image is only used to extract image-specific gradient in the user end with a compressed probe model, and the image-specific gradient is then fine-tuned in the server cloud with state-of-the-art model to generate the final adversarial noise. The end-cloud collaborated solution succeeds to resolve the paradox between the computation shortage in the end and the privacy leakage risk in the cloud. Moreover, to improve the attack performance of the generated adversarial perturbation in the cloud, we further introduce the image-independent universal adversarial perturbation for enhancement, which demonstrates effectiveness in both accelerated training convergence and stronger attacking capability.

The contributions of the paper can be summarized as follows:

- We designed an adversarial privacy-preserving filter to preserve users' portrait privacy from malicious face recognition crackers without affecting their photo sharing experience. Adversarial attack naturally meets the two basic requirements of *privacy* and *utility*.
- We proposed an end-cloud collaborated adversarial attack framework, which addresses the additional *non-accessibility* requirement to guarantee the original image only accessible to users' own device end. The compatible performance of this two-stage attack with the traditional one-stage attack (in Section 4.2.3) also provides a novel perspective to understand the adversarial attack problem.
- We integrated the universal adversarial perturbation with image-dependent perturbation to obtain improved privacy-preserving capability. This sheds light on alternative way to exploit adversarial examples.
- We conducted extensive experiments to validate the effectiveness and efficiency of the proposed solution framework. A prototype of adversarial privacy-preserving filter is further carried out and released for evaluation.

2 RELATED WORK

2.1 Privacy-preserving Photo Sharing

With the increasing popularity of Online Social Networks(OSNs), privacy-preserving photo sharing has received considerable attention [15, 28, 35]. Existing attempts can be roughly categorized into preserving image metadata [37], setting access control protocols [12] or sharing privacy policies [28, 29], and explicitly encrypting ROI region before uploading but decrypting after downloading [29, 35]. Different from these studies, the goal of this work is to hide the face identification information when sharing to OSN so that the potential image crawler and face cracker cannot use it for malicious usage.

2.2 Adversarial Attack

In the last few years, it has been witnessed that the existing machine learning models, not just deep neural networks, are vulnerable to adversarial attack [31]. Szegedy et al. [31] first introduced the problem of adversarial attack, and proposed a box-constrained L-BFGS method to find adversarial examples. To address the expensive computation, Goodfellow et al. [8] proposed Fast Gradient Sign Method (FGSM) to generate adversarial examples by performing a single gradient step, which later becomes a widely-used baseline attack method. Kurakin et al. [14] extended this method to an iterative version (I-FGSM). Dong et al. [5] further improved adversarial examples by adopting momentum term (MI-FGSM). Xie et al. [34] proposed DI^2-FGSM and M-DI^2-FGSM by adopting input diversity strategy, which focused on improving the transferability and achieved state-of-the-art performance in black-box setting.

There exist many attempts to explore the adversarial attack problem in face recognition applications. Sharif et al. [27] and Komkov et al. [13] proposed face recognition attacks by modifying facial attributes like adding virtual eyeglasses to impersonate other subjects or prevent them from being recognized. Dong et al. [6] proposed a query-based method for generating adversarial faces in black-box setting, which requires at least 1, 000 queries to the face recognition system. Some other works, e.g., AdvFaces [3] focused on employing Generative Adversarial Network (GAN) [7] to craft new adversarial images. However, although the above face recognition attack solutions easily satisfy the *privacy* requirement mentioned in Section 1, they may violate either the *utility* or *non-accessibility* requirement: (1) Attack solutions like modifying facial attributes and employing GAN tend to introduce non-trivial perturbation and make the perturbed face images unsuitable for sharing. (2) The above solutions all request to access the original face images, which leaves privacy-leakage risk in the cloud. Different from the existing face recognition attack solutions, in this work, we propose an end-cloud collaborated adversarial attack solution to simultaneously satisfy the *privacy, utility* and *non-accessibility* requirements. Moreover, the solution is compatible to allow instantiation with most of the existing adversarial attack algorithms.

In addition to the above image-dependent adversarial attacks, Moosavi-Dezfooli et al. [20] found a type of image-independent noise called universal adversarial perturbation (UAP), which can mislead the pre-trained model for different images. Following this, Poursaeed et al. [22] trained a generative network for generating universal adversarial perturbations (GAP), and Li et al. [16] observed the property of regional homogeneity and generated regionally homogeneous perturbations (RHP). These studies demonstrate the possibility of attacking the classification of specific images with universal perturbation. Inspired by this, on the basis of image-specific perturbations, we further enhance the adversarial examples

Table 1: Notation and explanations.

Symbol	Notation
$\mathbf{x}$	original image
$\mathbf{x_e}$	enrolled image
f_θ	face recognition model with parameter θ
g	naive gradient generated on probe model
$\tilde{g}$	gradient generated on server model
$\hat{g}$	transferred gradient
u	universal perturbation
$\hat{u}$	enhanced universal perturbation
s	enhanced adversarial noise
ϵ	maximum change restriction of each pixel
d	distance function
L	loss function
T	gradient transfer module
E	enhancement module
APF_g	adversarial image based on naive gradient
$APF\_\hat{g}$	adversarial image based on transferred gradient
APF_s	adversarial image based on enhanced adversarial noise
$Images\_u$	adversarial image based on universal perturbation

with universal adversarial perturbations in the cloud, with improved training convergence as well as attack performance verified in the later experiments.

3 METHODOLOGY

3.1 Problem Definition and Notations

Definition 1 (Adversarial Privacy-preserving Filter).

Given an original image $\mathbf{x}$, we assume that a face recognition model f_θ outputs an l-dimension vector as embedding feature $f_\theta(\mathbf{x})$. $\mathbf{x_e}$ represents the corresponding enrolled image, when the distance $d(f_\theta(x), f_\theta(x_e))$ between the two images is less than a certain threshold, they are judged to be the same subject, otherwise they are different subjects.

The main focus of this paper is to design the *adversarial privacy-preserving filter* (APF) that is an adversarial invisible noise s, making the adversarial example $\mathbf{x}+s$ look the same with $\mathbf{x}$ from the perspective of human observation but does not belong to the same subject from the perspective of face recognition model f_θ. Under the condition of ensuring image quality, it preserves usersâĂŹ portrait privacy without affecting their photo sharing experience.

The main notations of this work are summarized in Table 1.

3.2 Overall Framework

The working flow of privacy-preserving filter is illustrated in Fig. 2, where three parts are involved as user end, privacy-preserving server cloud and photo sharing service. Given a portrait image, its image-specific gradient is first extracted in the user end, and then issued to the privacy-preserving server cloud for enhancement to derive adversarial noise. After perturbed by the enhanced adversarial noise, the resultant adversarial portrait image is expected to fool the potential face cracker on photo sharing service. It is noted that the original image is accessible only to user end during the whole process, preventing information leakage even on the privacy-preserving server.

The core solution is an end-cloud collaborated adversarial attack framework, consisting of three algorithm modules deployed respectively in the user end and server cloud: (1) *Image-specific gradient generation* is to obtain the image-specific gradient by employing a compressed probe model runnable in the end. (2) *Adversarial gradient transfer* module is to align the image gradient from probe model and the server model in the cloud, with the goal to recover the adversarial information fooling practical crackers. (3) *Universal adversarial perturbation enhancement* module is to append image-independent universal perturbation to further enhance the derived adversarial noise. The following will elaborate each module.

3.3 Image-specific Gradient Generation

Following the *non-accessibility* requirement, users' original face images should not be uploaded to the cloud server to avoid leakage. However, most of the existing adversarial attack algorithms are too complicated to be deployed in the end due to their high demands on computing resources. Therefore, our solution is to employ a compressed model in the user end to extract preliminary information from original image to mismatch the enrolled image, and then enhance the information in the cloud.

Specifically, we use the compressed model as probe model in the user end to extract image-specific adversarial gradient g. Noted that different adversarial attack algorithms are allowed, which lead to g with different levels of intensity under the same maximum perturbation. In experiments we will discuss the influence of different adversarial gradient extraction algorithms. In this subsection, we employ FGSM as the example algorithm to introduce this process. Following standard adversarial attack operation, image-specific gradient g for original image $\mathbf{x}$ is extracted by:

$$g = \epsilon \cdot sign(\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{x_e}; \theta)) \tag{1}$$

where $\mathbf{x_e}$ is the corresponding enrolled image, θ denotes the network parameters, ϵ ensures that the generated gradient is within the ϵ-ball in the L_∞ space, and sign$(\cdot)$ denotes the sign function. $L(\cdot)$ is the loss function that measures the distance between the original face image and the enrolled image:

$$L(\mathbf{x}, \mathbf{x_e}; \theta) = -d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x_e})) \tag{2}$$

where $f_\theta(\cdot)$ represents the embedding feature, $d(\cdot,\cdot)$ is distance function. We use Euclidian distance in our work.

3.4 Adversarial Gradient Transfer

The small-scale probe model extracts the image gradient in the user end for generating adversarial samples. However, since the structure and parameters of probe model in the end and server model in the cloud are different, the gradient cannot be used directly. To align the image gradient between probe model and server model, we propose the gradient transfer module T, which is defined as:

$$\hat{g} = T(g) \tag{3}$$

We consider the gradient transfer module as an image-to-image translation network. The U-Net architecture is widely used in image-to-image translation problem, which is an encoder-decoder network

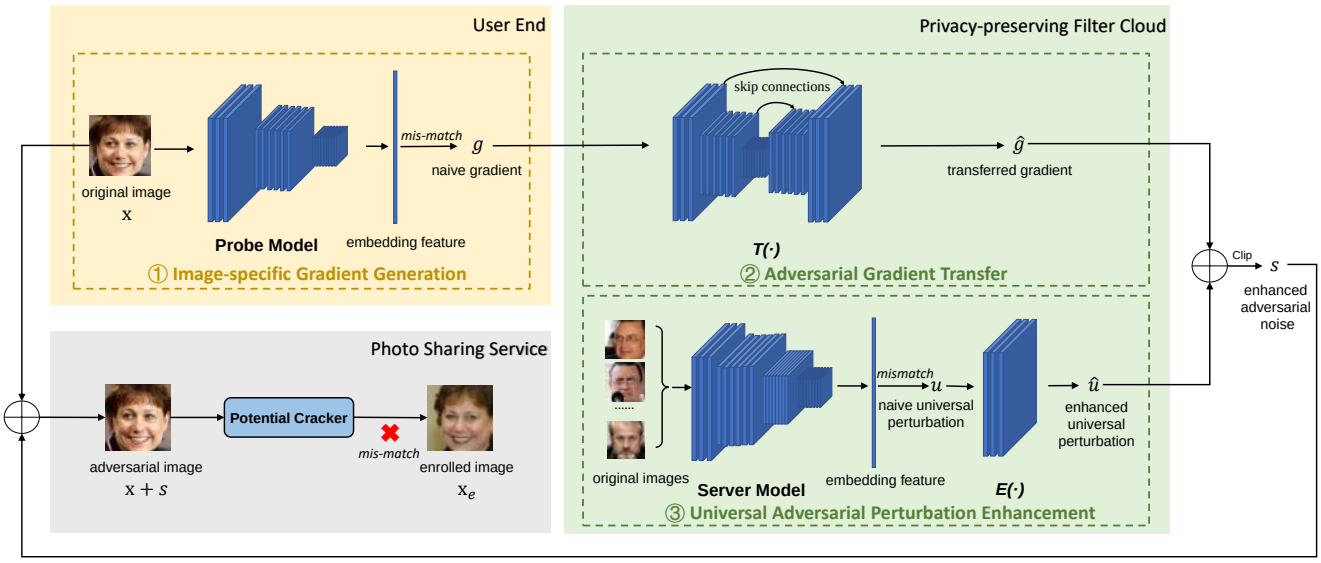

Figure 2: The proposed adversarial privacy-preserving filter framework.

with skip connections between the encoder and the decoder [23]. We employ this architecture for transferring gradients from probe model to server model. Specifically, we train the gradient transfer module as the following optimization problem:

$$\min_{\theta_T} \|\hat{g} - \tilde{g}\|_2 \tag{4}$$

where $\tilde{g}$ is the gradient generated on server model following Eqn. (1) and θ_T represents the parameters of network T. This objective function enforces the network T to learn the correlation between the probe model and server model.

Potential crackers may use multiple face recognition models, which is essentially a black-box adversarial attack problem. It is recognized that adversarial perturbation is transferable between models: if an adversarial image remains effective for multiple models, it is more likely to transfer to other models as well [18]. Inspired by this, to improve the attack performance to unknown cracking models, we implement the gradient transfer module not only for one-to-one domain transfer but also for one-to-many.

Specifically, given K white-box face recognition models with their corresponding gradient $\tilde{g}_1$, ..., $\tilde{g}_k$, we re-formulate the loss function in Eqn. (4) by replacing $\tilde{g}$ with $\tilde{g}_{ensemble}$:

$$\tilde{g}_{ensemble} = \sum_{k=1}^{K} \alpha_k \tilde{g}_k \tag{5}$$

where α_k is the ensemble weight with $\sum_{k=1}^{K} \alpha_k = 1$. For many face recognition models, the larger the value of K, the stronger the generalization capability of the derived adversarial images. However, an excessive K value will lead to high computational complexity and trivial weight α_k to underemphasize single model. We select $K = 2$ and evenly set $\alpha_k = 1/2$ [2]. In this study, we respectively select two of three state-of-the-art face recognition models as ensemble models for training and the remaining one as black-box model for testing. The performance of employing ensemble adversarial training to resist different face recognition models is reported in Section 4.3.1.

3.5 Universal Adversarial Perturbation Enhancement

As claimed in previous studies [16, 20, 22], universal adversarial perturbation contains image-independent information to mislead the classification of multiple images. This inspires us to integrate the image-specific information and image-independent information to enhance the performance of adversarial perturbation. Referring to the previous study [20], universal adversarial perturbation is a vector u that causes label change for most images sampled from the data distribution η:

$$\hat{C}(\mathbf{x} + u) \neq \hat{C}(\mathbf{x}) \quad \text{for most } \mathbf{x} \sim \eta \tag{6}$$

where $\hat{C}$ is a classification function that outputs for each image $\mathbf{x}$ an estimated label $\hat{C}(\mathbf{x})$. The previous universal adversarial perturbation studies produce a universal perturbation with the goal to cause image misclassification [16, 20, 22]. Instead of directly misleading image classification, we expect such a universal adversarial perturbation u, that provides a fixed image-independent perturbation deviating the examined face image far away from other face images. Therefore, we derive the naive universal perturbation as follow:

$$u = \max_{u} \sum_{i=1}^{n} \sum_{j=1}^{n} d(f_\theta(\mathbf{x}_i + u), f_\theta(\mathbf{x}_j)) \tag{7}$$

[2] Usually $\alpha_k = 1/K$ except when prior is available to emphasize on some of the models.

where n represents the number of all images contained in the dataset.

To integrate the image-specific perturbation and image-independent perturbation, we further design an enhancement module to adapt the domain of the universal perturbation u with the domain of $\hat{g}$. The enhancement module E consists of a scale transformation layer and a 1×1 convolution layer, which is defined as follows:

$$\hat{u} = E(u) = \text{conv}(\beta \cdot u + \gamma) \tag{8}$$

where β and γ are trainable parameters and $\text{conv}(\cdot)$ is the 1×1 convolution layer. The overall optimization problem incorporating the two proposed modules is as follows:

$$\max_{\theta_T, \theta_E} \; d(f_\theta(\mathbf{x} + \text{clip}_\epsilon(\hat{g} + \hat{u})), f_\theta(\mathbf{x}_e)) \tag{9}$$

where $\text{clip}_\epsilon(\cdot)$ indicates clipping the input within the ϵ-ball, θ_T denotes the parameters of network T, and θ_E denotes the parameters of network E.

Specifically, to employ ensemble adversarial attack, similar to Eqn. (5), given K face recognition models with their corresponding embedding feature $f_1, ..., f_k$, we reformulate Eqn. (9) to derive the final objective function by replacing $f_\theta(\cdot)$ with $\tilde{f}(\cdot)$ defined as follows:

$$\tilde{f}(\cdot) = \sum_{k=1}^{K} \alpha_k f_k(\cdot) \tag{10}$$

Network E and T are then trained by solving the above final objective function. The output $\hat{u}$ and $\hat{g}$ of E and T are added to derive s. The cloud will output s, and then return to user end to add it to the original image to derive the privacy preserved image, which can be safely shared to OSNs.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. We train the gradient transfer module T and the enhancement module E on the subset of MS-Celeb-1M dataset [9]. LFW dataset [11], AgeDB-30 dataset [21] and CFP-FP dataset [26] are used as test datasets to verify the effectiveness of the proposed privacy-preserving filter. We aim to jam the malicious face recognition algorithm to match the same users, thus the positive pairs (belong to the same person) are used for testing. In the experiments, we select subjects with two face images, where one is used as the enrolled image, and the other is for synthesizing the adversarial image.

- **MS-Celeb-1M** contains 10 million images. We randomly select 50,000 face images of 1,000 subjects for training, where each subject contains 50 face images.
- **LFW** contains 13,233 images of 5,749 different subjects. According to the refined version of Deng et al. [4], we use 6,000 images to construct 3,000 positive pairs of images.
- **AgeDB-30** contains 16,488 images of 568 different subjects. Same as above, we use 6,000 images to construct 3,000 positive pairs of images.
- **CFP-FP** contains 7,000 images of 500 different subjects. Same as above, we use 7,000 images to construct 3,500 positive pairs of images.

Face Recognition Models. In this study, we employ 4 state-of-the-art face recognition models to verify the effectiveness of the adversarial examples generated by our privacy-preserving models. MobileFaceNet [1] is served as probe model, and ArcFace [4] is used as default server model. To demonstrate the resistance to unknown cracking models, we introduce ArcFace, FaceNet [25] and SphereFace [17] as server models in Section 4.3.1.

- **MobileFaceNet** is specifically tailored for high-accuracy real-time face verification on mobile devices with 5.20MB model size, whose backbone network is MobileNet-V2 [24]. It was trained on MS-Celeb-1M dataset.
- **ArcFace** is the best public Face ID system, whose backbone network is Resnet-v2-152 [10]. It was trained on MS-Celeb-1M dataset.
- **FaceNet** and **SphereFace** were trained on CASIA-WebFace dataset [36]. The backbone networks of FaceNet and SphereFace are Inception-Resnet-v1 [30] and Sphere20 [17], respectively.

Evaluation Metrics. Recalling that the positioned privacy-preserving filtering problem is to both invalid the potential face recognition cracker and retain the original image quality, we introduce evaluation metrics regarding these two goals respectively.

For invaliding crackers, we use the standard attack success rate (ASR) [34] as the evaluation metric.

- **ASR** measures the effectiveness of our adversarial perturbation:

$$\text{ASR} = \frac{N_{w/o} - N_{w/}}{N_{total}} \tag{11}$$

  where $N_{w/o}$ and $N_{w/}$ denote the number of correctly recognized face images without and with perturbation[3], respectively. N_{total} denotes the total number of face images. The higher the ASR value, the better the adversarial perturbation effect, and the more satisfied the *privacy* requirement.

For retaining the original image quality, we quantify the similarity between the perturbed images and the original images via structural similarity (SSIM) [32].

- **SSIM** is a normalized metric whose values range from -1 to 1, which means the similarity from completely different image pairs to identical image pairs:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{12}$$

  Here, x and y are the two images to be compared, μ_x, μ_y, σ_x, σ_y, σ_{xy} are the means and variances of x and y, the covariance of x and y, respectively. $c_i = (d_i J)^2$, where $J = (2^{(\sharp \text{ of bits per pixel})} - 1)$, $d_1 = 0.01$, $d_2 = 0.03$ by default [33].

Implementation Details. We adopt ADAM optimizer with a fixed learning rate of 0.0001 for the entire network. Each mini-batch consists of 100 face images. The maximum perturbation ϵ of each pixel is set to be 8, which is more imperceptible for human observers

[3] In this study, we use L_∞ **distance** to restrict each pixel's change within a maximum scale, but with no limit on the number of pixels that are modified.

Table 2: The ASRs on LFW, AgeDB-30 and CFP-FP datasets.

Datasets	Adversarial Images	FGSM	I-FGSM	MI-FGSM	DI^2-FGSM	M-DI^2-FGSM	UAP	GAP	RHP
LFW	*APF_g*	74.5%	86.8%	88.2%	92.4%	89.1%	-	-	-
	APF_ĝ	91.9%	95.4%	89.8%	96.9%	96.5%	-	-	
	APF_s	94.8%	97.4%	95.7%	**98.8%**	**98.8%**	-	-	
	Images_u	-	-	-	-	-	27.9%	11.3	4.3%
	original image accessible	98.5%	99.4%	99.4%	99.4%	99.27%	-	-	-
AgeDB-30	*APF_g*	81.7%	86.3%	88.1%	90.5%	88.6%	-	-	-
	APF_ĝ	82.3%	90.8%	90.8%	94.9%	92.8%	-	-	-
	APF_s	88.3%	93.4%	93.8%	**95.5%**	94.7%	-	-	-
	Images_u	-	-	-	-	-	22.0%	23.7	13.1%
	original image accessible	95.8%	96.0%	96.0%	96.0%	96.0%	-	-	-
CFP-FP	*APF_g*	48.6%	57.2%	63.8%	68.3%	65.0%	-	-	-
	APF_ĝ	51.8%	72.8%	74.9%	84.7%	78.1%	-	-	-
	APF_s	67.4%	79.6%	82.8%	**88.3%**	85.3%	-	-	-
	Images_u	-	-	-	-	-	8.0%	6.1%	3.1%
	original image accessible	92.5%	93.7%	90.8%	93.2%	93.4%	-	-	-

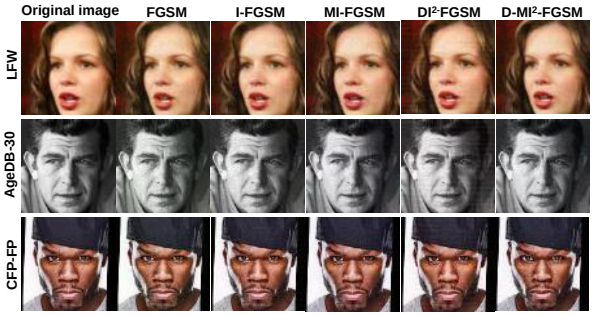

Figure 3: The example adversarial images perturbed by s. Three rows correspond to three different datasets, and each column corresponds to a different gradient extraction algorithm in Table 2.

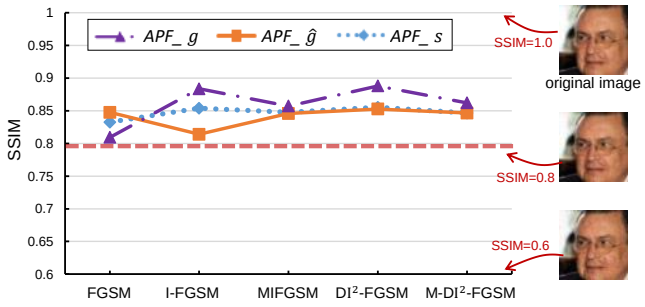

Figure 4: The SSIMs between the generated adversarial images and original images. The reference line 0.8 indicates the high quality of the adversarial images.

than general adversarial perturbation [19]. The embedding feature dimension l of MobileFaceNet, ArcFace, FaceNet and SphereFace is set to 192, 512, 128 and 512, respectively. Different face recognition models have different input sizes. We resize the input images to $112 \times 112 \times 3$ for ArcFace and MobileFaceNet, $160 \times 160 \times 3$ for FaceNet and $96 \times 96 \times 3$ for SphereFace. For adversarial gradient extraction algorithms (I-FGSM, MI-FGSM, DI^2-FGSM and M-DI^2-FGSM) are as follows: (1) For stochastic image transformations, we consider 3 transformations: rescaling, rotation and color conversion. (2) The probability of transformations is set to be 0.5. (3) The step size is set to be 2. (4) The total iteration number is set to be 30. (5) The decay factor of momentum term is set to be 1.

4.2 Performance Evaluation

*4.2.1 On **Privacy** and **Utility**.* Traditional adversarial attack methods are usually under the assumption of being accessible to original images. In this subsection, on the premise of satisfying the *non-accessibility* requirement, we discuss the experimental results of our end-cloud collaborated solution on different adversarial gradient extraction algorithms and datasets.

To demonstrate how our method meets the *privacy* requirement, the ASRs of all the compared methods based on 5 adversarial gradient extraction algorithms (FGSM, I-FGSM, MI-FGSM, DI^2-FGSM and M-DI^2-FGSM) are examined and shown in Table 2. *APF_g*, *APF_ĝ* and *APF_s* indicate the output adversarial images at three processing stages: generating images only based on the naive adversarial gradient g extracted by probe model in the user end, generating images based on the transferred gradient $\hat{g}$ in the cloud, and generating images based on the enhanced adversarial noise s which contains image-specific noise and image-independent universal perturbation. We have the following main findings: (1) In the case of meeting *non-accessibility* requirement, the ASRs of our proposed

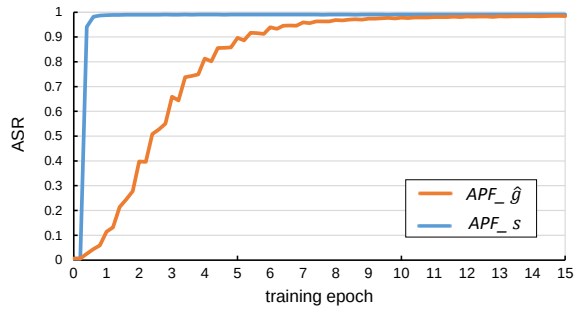

Figure 5: Convergence curve w/ and w/o universal perturbation enhancement.

method on the three datasets reach 98.8%, 95.5% and 88.3%, respectively. The good attack performance proves that the proposed filter can preserve user's *privacy* by fooling face recognition models. (2) Our privacy-preserving filter is compatible with various gradient extraction algorithms. The stronger the gradient extraction algorithms (viewing the results of the five algorithms shown in Table 2 from left to right), the better performance the filter achieves. This suggests us to use more effective adversarial gradient extraction algorithm when deploying the proposed framework in practical applications. (3) For each gradient extraction algorithm, *APF_ĝ* performs better than *APF_g* and *APF_s* achieves the best performance, which verify the effectiveness of the gradient transfer module and universal adversarial perturbation module.

To demonstrate how our method meets the *utility* requirement, some example images corresponding to *APF_s* mentioned in Table 2 are shown in Fig. 3. We can see that it is hard to tell the difference between original images and synthesized adversarial images from the perspective of human observation. We further use the quantitative metrics to measure image quality. SSIMs between the processed images and the original images on LFW dataset are illustrated in Fig. 4. We can observe that, when SSIM value of an image is greater than 0.6, the human eye can hardly tell the difference. The SSIMs of our processed images are all greater than 0.8, which indicates that the adversarial examples generated by our method can not only effectively interfere with the crack of face recognition, but also provide users with high-quality images to upload to the OSNs for use.

4.2.2 The Influence of Universal Adversarial Perturbation Enhancement. As mentioned in Section 4.2.1, universal perturbation can enhance the performance of adversarial images. In this subsection, we report further experimental results to examine the influence of universal perturbation enhancement. Testing ASRs of *APF_ĝ* and *APF_s* based on DI^2-FGSM of LFW dataset are shown in Fig. 5 with respect to the training epochs. We can observe that: (1) The ASR curve of *APF_s* is almost always above *APF_ĝ*, which is consistent with the observation from Table 2 that *APF_s* outperforms *APF_ĝ*. (2) *APF_s* converges nearly within an epoch, which is much faster than *APF_ĝ*. To understand the mechanism behind, we analyze

Table 3: Black-box ASRs on different training-testing pairs.

Training Models	Testing Models			
	FaceNet	SphereFace	ArcFace	Average
raw	83.0%	50.2%	92.5%	75.1%
FaceNet	95.2%	75.6%	98.3%	89.6%
SphereFace	91.5%	85.6%	96.7%	91.1%
ArcFace	93.8%	73.3%	98.8%	88.5%
Ensemble(F+S)	95.1%	81.8%	**98.6%**	**91.7%**

the training process of the network. For ease of description, the loss functions in Eqn. (9) are defined as L_s. Universal adversarial perturbation u will be transmitted to L_s by forward propagation, then $\nabla_{\theta_T} L_s$ will affect the training of gradient transfer module by backward propagation. The enhancement module itself is easy to train due to very few parameters it has. So the convergence speed is accelerated.

Moreover, three kinds of adversarial images *Images_u* perturbed by only the universal adversarial attack methods (UAP, GAP, RHP) are also implemented for comparison. As shown in Table 2, we can observe that these universal adversarial attack methods fail to achieve a reasonable ASR. This demonstrates that, while universal adversarial perturbation has potential to simultaneously attack multiple images, it singly cannot guarantee a promised attack performance compared with traditional attack methods utilizing image-specific information.

We propose a strategy that combining universal adversarial perturbation with image-dependent adversarial gradient, and achieve a productive performance. We hope this study could draw attention to this combination on the other adversarial attack tasks, e.g., classification, object detection, semantic segmentation.

*4.2.3 On **Non-accessibility**.* To examine the potential of the proposed two-stage attack solution to approximate the traditionally generated adversarial perturbation, we further compare APF with the adversarial attack setting violating the *non-accessibility* requirement, noted as *original image accessible* in Table 3. For this setting, the original images are assumed to be exposed to the server, where adversarial images are generated by 5 adversarial gradient extraction server models. As shown in Table 3, the proposed end-cloud collaborated privacy-preserving solution (*APF_s*) obtains comparable performance with the setting accessible to original images. This demonstrates that, even without directly accessing the original images, it is possible to achieve remarkable attack performance by conducting auxiliary operations such as gradient transfer and universal perturbation enhancement introduced in this study. This result on one hand opens up possibility for alternative way of implementing adversarial attack, but on the other hand imposes the new challenge to adversarial defense in the future when original samples are not available.

4.3 Real-world Implementation Discussion

4.3.1 Transferability and Black-box Attack. Since what face recognition models the cracker use is unknown and typical crackers

Table 4: ASRs for *adversarial-original* and *adversarial-adversarial* matching.

		Adversarial-original	*Adversarial-adversarial*
DI^2	w/o δ	90.2%	**40.6**%
	w/ δ	86.7%	96.4%
M-DI^2	w/o δ	98.3%	**27.2**%
	w/ δ	98.4%	94.7%

may use multiple face recognition models, in practice the privacy-preserving effectiveness essentially depends on the solutionâĂŹs generalization and transferability under black-box attack settings. To address this, we implemented 3 server models of ArcFace, FaceNet and SphereFace, and evaluated the attack performance on *APF_s*. The adversarial gradient extraction algorithm is fixed as DI^2-FGSM, and the ASRs are calculated based on the LFW dataset. To construct the black-box attack setting, among the 3 server models, FaceNet and SphereFace are selected as white-box models, and ArcFace is selected as the black-box model due to its wide utilization in public Face ID system[4]. In addition, we also examined the performance of the ensemble adversarial attack combining FaceNet (F) and SphereFace (S).

Table 3 shows the black-box ASR under different training-testing pairs. For example, the value of 98.6% represents the ASR trained with ensembled white-box models and tested on ArcFace. Higher ASR value means superior resistance performance to malicious face recognition model and better transferability of the method. There are several observations: (1) The adversarial images (the last four lines) generated by any model outperform than these adversarial images (first row) generated by naive adversarial gradient extraction algorithms. This demonstrates our proposed framework can improve the transferability of adversarial images, without limitation with specific models. (2) The adversarial images generated with a special model perform well when they are tested by the corresponding model, but slightly worse on the other models. It is expected that the white-box attack has better performance than black-box attack does. (3) When the adversarial images are generated by ensemble models and are attacked by ArcFace model (as the black-box), the ASR (98.6%) is higher than the other two black-box methods (95.1%, 81.8%), even almost catches up with the white-box ASR (98.8%). It verifies the transferability of our proposed method in employing ensemble training towards black-box attacking. (4) The average ASRs of ensemble training model is the highest among all training models. It is expected with more models implemented in ensemble training, the ASR performance towards arbitrary black-box attacking methods will be guaranteed. Referring to previous the study [18], it is reasonable to choose the model with the large structure differences to employ ensemble adversarial training. In practical applications, we can carefully select widely-used white-box models with typically different structures to improve the generalization and transferability to specific models.

[4] The black-box model is to simulate the possible cracking face recognition choices in real-world applications. Therefore, a widely employed model can better evaluate the privacy-preserving performance in practice.

4.3.2 Adversarial-adversarial Image Matching. The above experiments all assume that the enrolled image is original image without adversarial perturbation. However, in practice, there is a chance when the enrolled image collected by face crackers is already adversarially filtered, i.e., the problem turns to match between the adversarial images from unique user. To examine the performance of the proposed solution under this situation, we randomly selected 5,000 testing images from 100 subjects in the MS-Celeb-1M. Among the 50 images for each subject, we set one image as the enrolled image and the remaining 49 to generate the adversarial images.

Following Eqn. (1) and (2) to generate adversarial images, we evaluate the following two settings for each subject: (1) *adversarial-original*, matching between the 49 adversarial images and the original enrolled image; (2) *adversarial-adversarial*, matching among the 49 adversarial images. The resultant average ASR is reported in Table 4 using adversarial attack methods of DI^2-FGSM and M-DI^2-FGSM (first row for each method). The results show that it is easy to match between two adversarial images for the same subject (with ASR of 40.6% and 27.2%) and the proposed solution fails to attack the face cracker under this situation. We owe this result to that with the fixed feature of enrolled image $f_\theta(\mathbf{x_e})$ in Eqn. (2), the generated adversarial images from the 49 original images tend to be similar and are easy to be grouped into one subject. Therefore, we introduce a random noise δ to modify Eqn. (2) as follows:

$$L(\mathbf{x}, \mathbf{x_e}; \theta) = -d(f_\theta(\mathbf{x}), f_\theta(\mathbf{x_e}) + \delta) \tag{13}$$

where δ is a random vector with 10% elements following a uniform distribution $U(-0.07, 0.07)$ and the remaining 90% setting as 0. δ acts as a random rotation deviation from the feature of the enrolled image, so to prevent the generated adversarial images overfitting along unique gradient direction. The results of generating adversarial examples via Eqn. (13) are shown in the second row for each attack method in Table 4. It is observed that the introduction of δ significantly improves ASR for *adversarial-adversarial* while basically maintains the high ASR for *adversarial-original*. This modification guarantees the privacy-preserving effectiveness of our solution.

5 CONCLUSION

In this study, we introduce a portrait photo privacy-preserving solution when sharing to OSNs to resist malicious crawler and face recognition usage. The proposed end-cloud collaborated adversarial attack solution is validated to well satisfy all three requirements of *privacy*, *utility* and *non-accessibility*.

In the future, in addition to testing the solutionâĂŹs effectiveness in practical trials, we are also interested to work towards the following two directions: (1) the combination of image-specific and image-independent perturbation in more attack scenarios, (2) the attack and defense attempts when the original sample is not accessible.

ACKNOWLEDGMENTS

This work is supported by the National Key R&D Program of China (Grant No. 2018AAA0100604), and the National Natural Science Foundation of China (Grant No. 61632004, 61832002, 61672518, U19B2039, 61632006, 61772048, 61672071, and U1811463), and the

Beijing Talents Project (2017A24), and the Beijing Outstanding Young Scientists Projects (BJJWZYJH01201910005018).

REFERENCES

[1] 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*. Springer, 428–438.

[2] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107 (2019), 1753.

[3] Debayan Deb, Jianbang Zhang, and Anil K Jain. 2019. Advfaces: Adversarial face synthesis. *arXiv preprint arXiv:1908.05008* (2019).

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4690–4699.

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9185–9193.

[6] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. 2019. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7714–7722.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*. 2672–2680.

[8] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[9] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 87–102.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 630–645.

[11] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*.

[12] Peter Klemperer, Yuan Liang, Michelle Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael Reiter. 2012. Tag, you can see it! Using tags for access control in photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 377–386.

[13] Stepan Komkov, Aleksandr Petiushko, and al et. 2019. AdvHat: Real-world adversarial attack on ArcFace Face ID system. *arXiv preprint arXiv:1908.08705* (2019).

[14] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2016. Adversarial examples in the physical world. *Proceedings of the International Conference on Learning Representations (ICLR) workshop*.

[15] Fenghua Li, Zhe Sun, Ben Niu, Yunchuan Guo, and Ziwen Liu. 2018. Srim scheme: An impression-management scheme for privacy-aware photo-sharing users. *Engineering* 4, 1 (2018), 85–93.

[16] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan L Yuille. 2019. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses. *arXiv preprint arXiv:1904.00979* (2019).

[17] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 212–220.

[18] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. 2017. Delving into Transferable Adversarial Examples and Black-box Attacks. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[19] Yan Luo, Xavier Boix, Gemma Roig, Tomaso Poggio, and Qi Zhao. 2015. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292* (2015).

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1765–1773.

[21] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. 2017. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 51–59.

[22] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. 2018. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4422–4431.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 234–241.

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4510–4520.

[25] 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 (CVPR)*. 815–823.

[26] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. 2016. Frontal to profile face verification in the wild. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–9.

[27] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security*. 1528–1540.

[28] Jose M Such and Natalia Criado. 2016. Resolving multi-party privacy conflicts in social media. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1851–1863.

[29] Weiwei Sun, Jiantao Zhou, Shuyuan Zhu, and Yuan Yan Tang. 2018. Robust privacy-preserving image sharing over online social networks (osns). *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1 (2018), 1–22.

[30] 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 AAAI Conference on Artificial Intelligence*.

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[32] 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 (TIP)* 13, 4 (2004), 600–612.

[33] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. 2003. Multiscale structural similarity for image quality assessment. In *Proceedings of the Asilomar Conference on Signals, Systems & Computers*, Vol. 2. 1398–1402.

[34] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2730–2739.

[35] Yi Xu, True Price, Jan-Michael Frahm, and Fabian Monrose. 2016. Virtual U: defeating face liveness detection by building virtual models from your public photos. In *Proceedings of the USENIX Conference on Security Symposium*. USENIX Association, 497–512.

[36] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[37] Lan Zhang, Kebin Liu, Xiang-Yang Li, Cihang Liu, Xuan Ding, and Yunhao Liu. 2016. Privacy-friendly photo capturing and sharing system. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 524–534.