Mathematical Biosciences and Engineering

MBE, 20(2): 1820–1840.
DOI: 10.3934/mbe.2023083
Received: 17 September 2022
Revised: 13 October 2022
Accepted: 24 October 2022
Published:7 November 2022

http://www.aimspress.com/journal/MBE

Research article

PrivacyMask: Real-world privacy protection in face ID systems

Guangmin Sun[1], Hao Wang[2,*], Yu Bai[3], Kun Zheng[1], Yanjun Zhang[2], Xiaoyong Li[2] and Jie Liu[1]

[1] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
[2] Network and Information Technology Center, Beijing University of Technology, Beijing 100124, China
[3] Medical Engineering Division, Beijing Friendship Hospital, Beijing 100050, China

* **Correspondence:** Email: wanghao@bjut.edu.cn; Tel: +8601067392193.

Abstract: Recent works have illustrated that many facial privacy protection methods are effective in specific face recognition algorithms. However, the COVID-19 pandemic has promoted the rapid innovation of face recognition algorithms for face occlusion, especially for the face wearing a mask. It is tricky to avoid being tracked by artificial intelligence only through ordinary props because many facial feature extractors can determine the ID only through a tiny local feature. Therefore, the ubiquitous high-precision camera makes privacy protection worrying. In this paper, we establish an attack method directed against liveness detection. A mask printed with a textured pattern is proposed, which can resist the face extractor optimized for face occlusion. We focus on studying the attack efficiency in adversarial patches mapping from two-dimensional to three-dimensional space. Specifically, we investigate a projection network for the mask structure. It can convert the patches to fit perfectly on the mask. Even if it is deformed, rotated and the lighting changes, it will reduce the recognition ability of the face extractor. The experimental results show that the proposed method can integrate multiple types of face recognition algorithms without significantly reducing the training performance. If we combine it with the static protection method, people can prevent face data from being collected.

Keywords: privacy protection; neural network; facial recognition; adversarial attacks; spatial mapping; transferability; nonlinear transformation

1. Introduction

Nowadays, cameras with face recognition algorithms (FRA) are everywhere. Not only do FRAs get our facial features, but they are likely not to tell us what they are doing. What is worse, even people wearing masks or sunglasses cannot avoid face detectors. Especially after the COVID-19 pandemic, almost all commercial FRAs support the identification of occluded faces. Therefore, our facial privacy is under an unprecedented threat. Also, we cannot eliminate the collection of personal biological data in the physical world.

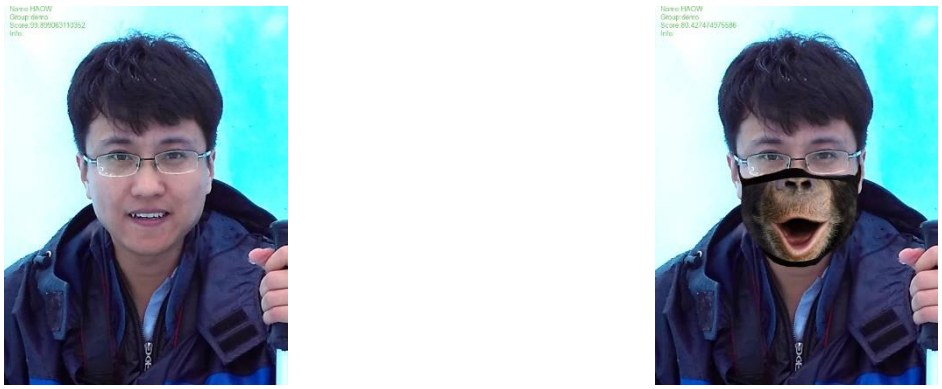
Figure 1. A CCTV that recognizes masks and face IDs. By examining the closed-circuit television video through mask identification, we can know how many people follow the health guidance in specific areas, such as the Little Italian neighborhoods in New York city, provided by Tryo labs.

In the past, people could easily hide their faces by wearing a mask, a hat or a pair of sunglasses. Some ingenious attack networks can generate printable patches [1] for an FRA and place them in parts of the body [2,3] to mislead the classifier [4]. Of course, numerous algorithms can protect static face privacy data. However, recent surveys have proved that our facial features can still be inferred when the faces don masks and other ornaments. As shown in Figure 1, CCTV in many neighborhoods can identify whether people are wearing masks and the state-of-the-art FRAs can still distinguish the face ID. These innovative recognition algorithms invalidate many liveness privacy protection algorithms, and most importantly, these models are freely available to anyone on the internet.

The existing occlusion face recognition methods include representation-based and reconstruction-based algorithms. The online evaluation uses a gradually growing network approach to remove an unwanted object. In [5], two discriminators are used to learn the global structure and the deep missing

region for face image reconstruction. Similarly, Ge et al. [6] applied a set of identity-centered features in the extractors as supervision to enable the clustering toward their identity centers. In [7], the authors restore the whole face with a local facial texture. In the reconstruction field, the deep features extracted from the unmasked regions are widely used [8]. Furthermore, the accuracy of the MTArcFace [9] algorithm to recognize the face donning a mask has exceeded 99.7%. To sum up, those recognition algorithms are very accurate for faces donning masks. In contrast, the development of facial privacy protection methods is much slower than that of FRAs.

Attacking the existing FRAs is an effective means to protect face privacy in the physical world. Face information is obtained mainly through two modes. One is the face features to be sampled into multiple blocks of the same size. And, the other matching approach is detecting the key points. According to these theories, some methods have used cloak texture to protect IDs in face images. In other words, people use printed textures with deceptive features to deceive the camera and place them on clothes [4], hats [2], eyeglass frames [10] and other parts [3] of the liveness body. This device can play a role in the early stage. Nevertheless, facial recognition systems try to protect themselves from being attacked. For example, they can determine human beings by analyzing the wavelength of reflected light [11]. Kose and Dugelay [12] use the method based on linear back projection to discriminate between real and fake faces, and they asserted 88% classification accuracy with a depth map. Face synthesis [13], deepfakes [14], attribute manipulation [15] and expression swap [16] are good ways to protect privacy. However, they may be easily exploited by criminal motives. Some new in-depth definition identification technologies have been proposed to deal with privacy protection methods, as shown in Figure 2. Therefore, liveness privacy protection becomes more difficult.



(a) (b)

Figure 2. Interference effect of a mask with a simulated pattern. (a) The recognition rate is 99.899% when there is no occlusion, (b) the interference pattern can reduce the recognition rate, but the face confidence is more than 80%.

To solve the optimization problem, various multi-objective evolutionary algorithms have been proposed. We can improve them by leveraging domain-specific information in training signals for related tasks. We proposed a privacy framework and showed that solving such a game is equivalent to solving a one-to-many objective linear programming problem. Its antagonism is a complex problem because there are multiple ways to sense the generated attack image, whereas many previous methods have failed. The improved training network makes the machine automatically perceive external

changes and adds a manual intervention mechanism. That can effectively improve the efficiency of the attack. All targets in mutual updating are black box algorithms. The key to this model is to preserve the trained parameters, accumulate new parameters at the inference level and maintain visibility at the image level as much as possible.

This paper focuses on protecting people's facial privacy in the real world. First, by studying the types and principles of FRAs, we prepare an attack method against commercial face detectors and improve their efficiency. We adopt the attention mechanism to keep the network stable and robust. Simultaneously, it is easy to fuse FRAs for multiple patterns. Second, we try to train some 2D images that can deceive any FRA. The difference from previous studies is that the protective texture can be compatible with most commercial FRAs. Finally, we design a projection network, which can project the attack texture from 2D to 3D. It will be easy to adapt the image to the mask. This 3D texture makes performance independent of angle and lighting. More importantly, these textures are natural from a visual point of view and will not make others feel suspicious and uncomfortable.

2. Related work

Protecting the private information in face images can be divided into two modes: one mode is by adding interference in the digital domain (digital attack), and another is by adding anti-disturbance to real faces in the physical world (physical attack). Because attackers generally cannot access or modify the digital domain input of the physical world face recognition system, physical attacks have more practical significance for evaluating the robustness. However, compared with digital attack methods [17], the physical attack is more challenging due to the complex conditions of the physical world.

2.1. Occlusion face recognition

We classify the principle of occluded face recognition of three types: facial reconstruction, occlusion discarding and occlusion dictionary [18]. A linear combination of gallery images rebuilt the probing facial picture in facial reconstruction. The occlusion discarding strategy seeks to delete characteristics in FRAs that correspond to the occluded component. It is on the notion that the features damaged by the occluded section have a detrimental influence on the FRA.

In reconstruction, Alyuz et al. [19] extended studies about occlusion removal and restoration techniques. They consider occlusion handling for surface registration and missing data handling for classification based on subspace analysis techniques. It is easy to use an adaptively selected model-based registration technique for the alignment problem. Occlusions are recognized and eliminated after registering to the model that has been validated by non-occluded patches. A masking approach known as masked projection is presented during the classification step to allow subspace analysis techniques with insufficient data. Drira et al. [20] attempted to expect and complete the incomplete face curves by using a statistical shape model. In particular, this technique employs the principal components analysis (PCA) algorithm on the training data to generate a training shape, which is then used to estimate and rebuild the entire curves. The approach of recovering occluded regions by predicting missing curves that incorporate critical form data dramatically improved identification accuracy.

Regarding occlusion discarding, Li et al. [21] presented a restoration step that employed a statistical estimate on curves to address missing data. In particular, the model created by PCA builds

partially observed curves. Guo et al. [22] used inverse rendering to make many photo-realistic facial pictures with varying features. They established a fine-detailed facial picture collection by transferring multiple detail scales from one image to the next.

Regarding the occlusion dictionary, Hong et al. [23], Utomo and Kusuma [24] and Prinosil and May [25] have realized person re-identification from different angles and produced certain benefits. Even commercial face identification software development kit (SDK) such as Baidu, Aliyun, Tencent, ArcFace (not InsightFace) and Face++ have made the accuracy of masked face recognition over 99%. At present, no matter how we cover up our facial privacy in the physical world, we can be captured by advanced FRAs. Tolosana et al. [26] have focused on the newest generation of deepfakes, emphasizing its advancements and problems for fake detection.

2.2. Living face privacy

Initiative protection mode is based on the detector's compliance with the privacy convention. FRAs actively conceal the privacy of the captured picture and transmit clean data through the network. Blind vision involves the anonymous processing of an image or video. It uses a secure multi-party computing (SMC) method applied to visual algorithms to solve the privacy problem during processing [26]. SMC is a subfield of cryptography that allows several people to compute a function while keeping their inputs and themselves hidden. Sun et al. [28] used dynamic keys for decryption during image transmission, thus avoiding the risk of disclosure of fixed keys. Zhou and Pun [29] developed a new method called face pixelation in video live streaming to generate automatic personal privacy filtering during unconstrained streaming activities. RGB cameras for active and assisted living applications [30] use object segmentation technology to process the images collected from the community in the back progress and hide the irrelevant face data in the pictures. The deepfake [14] mentioned above is also a way to protect privacy, although it is usually used for illegal purposes. Zheng et al. [31] used remote photo-plethysmography (RPPG) and region of interest with a high signal-to-noise ratio to forge physiological signals to protect the privacy of the subjects. Chen et al. provided evaluations of PulseEdit's [32] performance against RPPG-based liveness identification and RPPG-based deepfake detection, demonstrating its capacity to avoid these visual security techniques and its critical role in the design of attack-resilient systems. Fawkes algorithm proposed by Shan et al. [33] hid the small perturbation in the position where the image is difficult to be detected, resulting in the loss of focus on the face detector.

Passive protection has five categories: intervention, blind vision, secure processing, redaction and data hiding [34]. Although it often appears in static data protection, passive protection often comes with liveness data protection events in the physical world. The intervention approach addresses the problem of preventing someone from capturing private visual data from their surroundings. They intend to build environments that are resistant to capture. These methods interfere with the camera equipment by interfering with the special equipment of the camera's optical lens, thus hindering the acquisition of pictures. The BlindSpot system was created by Patel and colleagues, and it detects many retro-reflective CCD or CMOS camera lenses in a safe area and then focuses a pulsating light toward the identified lens [35], destroying any photos those cameras may have recorded. Adversarial patches [36], AdvHat [2] and federated privacy [3] offer real-world approaches to creating universal, resilient and targeted hostile picture patches. When these patches are placed with a living face, they can cause an FRA to ignore the face or report an error target.

3. Materials and methods

Physical attack methods generally include face stickers, wearing adversarial glasses, wearing a hat, etc. We should also consider different physical world conditions, including the chromatic aberration of a sticker, face variations and environmental condition variations. These are closely related to lighting conditions, camera angle, etc. Obviously, here are two methods to protect the privacy of physical faces: dodging attacks and impersonation attacks. The principle of dodging attacks is to reduce the similarity between face images with the same ID, which generally refers to cosine similarity ($L_{sim} = (1 + \cos f(x + \delta), f(x^a))/2$), where x and x^a come from the same ID and L_{sim} stands for adversary loss.

The principle of an impersonation attack is to increase the similarity between face images with different IDs. In contrast to avoiding attacks, x and x^a are sampled from two unequal identity IDs and can be optimized by minimizing $L_{sim} = (1 - \cos f(x + \delta), f(x^a))/2$.

3.1. *Algorithm fusion*

The principle of privacy protection is an attack recognition algorithm. The real meaning of privacy protection is algorithm fusion because the existing face privacy protection algorithms are specified recognition models. However, the face recognition extractors in the physical world are different. They may lead to failure in reality because of using an independent theory. Further, we cannot receive the practices and results consumed by the multi-deep network.

Because many precedents have described how to tease FRAs with image interferences, privacy attacks on multiple FRAs have become rule-based. If a gauze mask is printed with a texture with general attack attributes, people who wear it will look more natural. We use the encoder to map some opponent features and mask shapes so that these faces can be regarded as another living person by the FRAs. The encoder abstracts the eigenvalues of the synthesized image and inputs them into several spliced FRAs. Then, the decoder feeds back the classification results of these spliced FRAs to the generator to adjust the attack parameters. When the synthesized texture reaches the attack threshold, the parameters can be output to the printable network to make a mask texture. The structure of the fusion network is shown in Figure 3.

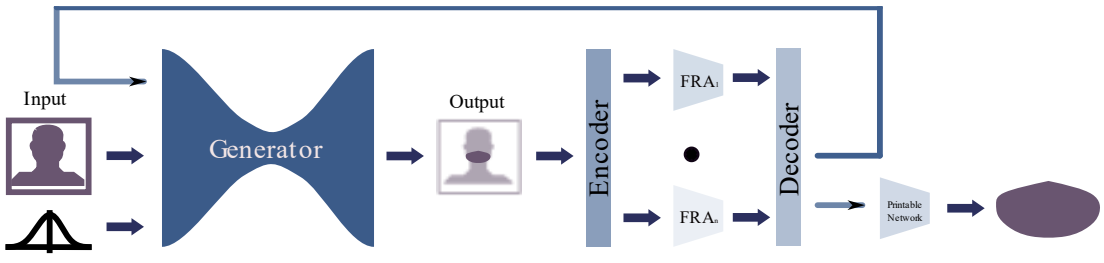
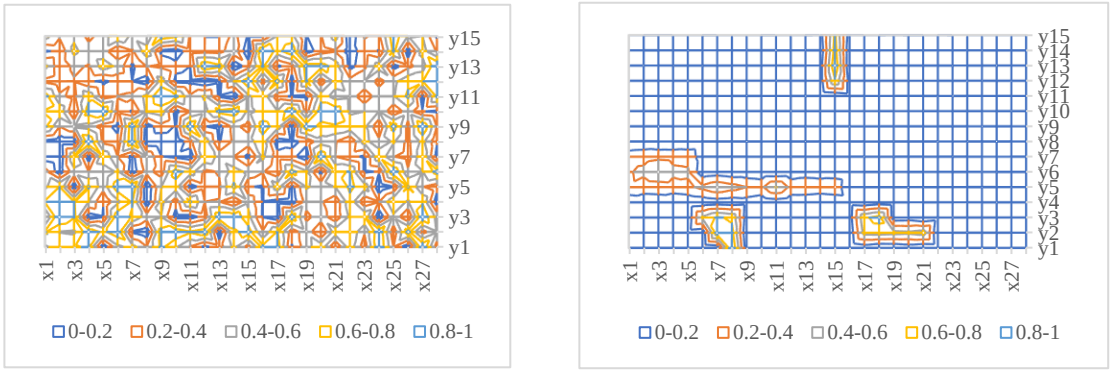


Figure 3. Structure of fusion network.

Machine learning, as everyone knows, has generalization performance that is poor, but it often produces more optimized results on the specific problem with sufficient training. So, keeping the model available when changing the input and output is a challenge. Some traditional stitching ideas

can be applied in deep networks to simulate the generalization process, e.g., wiring together a collection of known algorithms. This smoother representation of the algorithm allows learning to adjust the internal mechanism of the algorithm itself through data feedback.

The biggest problem we encountered by fusing face recognition attack algorithms is comprehending the internal concerns of a commercial algorithm. All black-box-based algorithms should be paralleled into the attack network to achieve compatibility. Although visual attention [37] technology is often used for character recognition to simulate the human eyes, similarly, it will bring surprising results in algorithm fusion. If a network can sense key nodes, its speed must be faster, and its accuracy will be higher than that of the pruning method. Squeezing eigenvalues of the same person's face into small intervals is a good way of increasing the convergence speed of a network, as shown in Figure 4. However, with the accumulation of recognition algorithms, the capacity of feature space needs to be continuously adjusted, and gradient explosion may also occur in theory. Therefore, we should use the weighting function to interfere with the aggregation speed. See Eq (1), where w is the ith feature weight, l is the length of the face feature vector, x represents the attention distribution coefficient in the ith feature and y is used as the actual eigenvector estimation. Equation (2) defines the probability distribution of x, where s is the attention score [38].



(a) (b)

Figure 4. Eigenvalue extrusion in encoder. (a) Eye eigenvalue distributions of different FRAs and (b) distribution extrusion of eye eigenvalues for different FRAs.

Unlike in natural language recognition [39], the attention mechanism of facial features is a multi-vector parallel query process, and the vector length is non-fixable. So, based on the first FRA, the filters are sorted according to the filter norm (e.g., L1 or L2) [40,41]. Such processing will reduce the resources occupied by the decoder so that the network optimization will not be slow because of the cask effect. The decoder normalizes the cut feature map and feeds it back to the generator, and the feature mapping set $\boldsymbol{D}$ represents the decoder, which is composed of multiple vectors ($\boldsymbol{A}$) with the weight w, as shown in Eq (3). On the one hand, the decoder outputs the recognition result of the polymerized algorithm. On the other hand, it feeds the area of interest back to the generator. When the texture structure generated by Gaussian noise reaches the preset threshold, it can automatically or manually convert vector data into 2D coordinates for printing. Moreover, the ideal result should be to shift the attention of any target recognition algorithm. The image based on the attention mechanism is shown in Figure 5.

$$w_i = \sum_{i=1}^{l} x_{ij} y_j \tag{1}$$

$$x_i = \text{softmax}(s_i) = \frac{\exp(s_i)}{\sum_j^l \exp(s_j)} \tag{2}$$

$$\boldsymbol{D} = \frac{w_1\boldsymbol{A_1} \oplus w_2\boldsymbol{A_2} \oplus \cdots \oplus w_n\boldsymbol{A_n}}{\|\boldsymbol{A}\|} \tag{3}$$

(a) (b) (c)

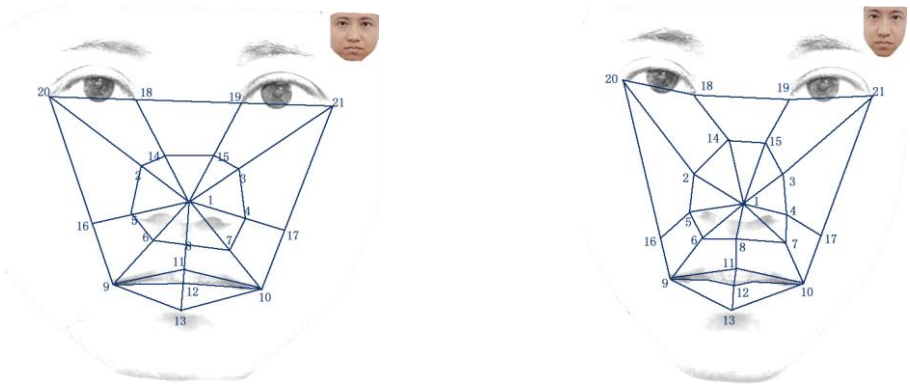
Figure 5. Fusion of FRAs based on an attention mechanism. (a) Generated mask on face model, (b) feature points of the two FRAs squeezed by the encoder and (c) attention shift model after decoding.

3.2. Texture projection

Although simulation stickers can use attacks to achieve the effect of privacy protection, considering the changes of light and angle in the wild world, it is easy to invalidate the planar texture pattern. Therefore, the generated texture must be projected onto the mask with three creases. The projection here includes two parts: one is the projection of the mask on the face, and the other is the projection of the texture on the mask. Further, we convert it from 3D to 2D, and then convert it to 3D again.

First, the discrete point elimination method [42] is used to record the 3D scattered point information of the human face in which the coordinate is p. We put the virtual mask onto a face image with the assistance of surface shape and curvature scales [43], where δ represents the network topology coefficient. To be sure, aligning the mask with the nose coordinates as the starting point makes the mapping between the 2D and the 3D plane more symmetrical. Locking the range of the geometric surface indices in [0.85,1] can make the image fit with an object surface as much as possible [44], especially when mapping the nasal tip area. For this training, we chose to use the FRGCv2 database [45]. The 2D space shall conform to Eq (4), where the sequence comprises the L2 norm maximum of p between UV space and XYZ space [46]. The face after transposition is closer to the tension on the plane, as shown in Figure 6.

$$d_{x,y} = \max_{\delta \in \{-1,0,1\}} \left\| p_{u,v} - p_{x+\delta,y+\delta,z+\delta} \right\| \tag{4}$$



(a) (b)

Figure 6. 2D expansion of a human face. (a) Front of 3D image and (b) 3D image tiled into a plane. Most facial features are planarized.

Second, the mask needs to be laid flat on the 2D face and passed through a suitable position. Usually, the size of masks that meet the biological evaluation of medical devices (according to the ISO10993-1 standard) is 17.5 cm × 15.5 cm after deployment. So, the outspread mask must be larger than the area under the nose. We chose to apply the Dirichlet boundary condition [47] and the energy E defined in Eq (5), where F_m symbolizes the function of linear transformation constrained by piecewise constraints, s_m represents the boundary of 3D to 2D mapping and μ is the unit slice of the 3D face surface. This convex mapping can constrain the discrete conformal graph (maximum angle hold). Therefore, it is only necessary to ensure that the nose and the first outer edge are mapped as shown in Figure 7. And, the area below the third inner edge can be ignored. So far, we can print the texture on the 2D mask according to the face shape.

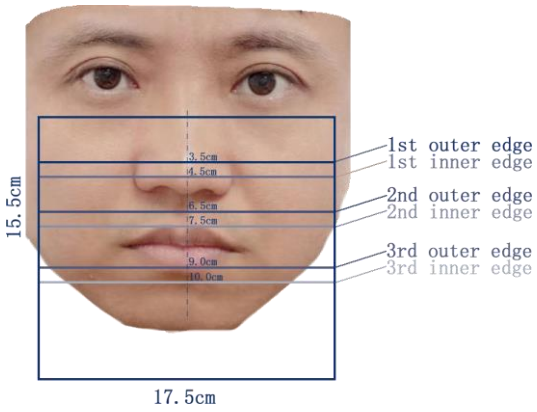


Figure 7. Conformal mapping of mask and 2D face. The mapping result can be represented by orthogonalizing the positive highlight area (nose tip) of the 2D image with the first outer edge of the mask. The nose tip is usually 3.5 cm away from the upper edge of the mask.

Finally, considering that the mask will block some texture information after being folded, a grid training network needs to predict which features will fail when they are occluded. For positioning, we print 1.5 cm chessboard grids on the prefabricated masks. It should be noted that most of the parts

cause large-scale texture distortion in blue areas, as shown in Figure 8. A 2D set $g_i = \{x_i^t, y_i^t\}$ denotes the pixels of a regular grid, and x_i^s presents the source pixel in the expanded mask, as shown in Figure 7. By the four points of the *i*th cell and corresponding projected cell, it can be transformed to the matrix of the spatial M_θ, as shown in Eq (6). The class of transformations $\mathcal{T}_\theta$ may be more constrained, such as that used for attention, allowing cropping, translation and isotropic scaling [46].

$$E = \frac{1}{2}\int s_m |\nabla F_m| d_\mu \tag{5}$$

$$(x_i^s, y_i^s) = \mathcal{T}_\theta(x_i^t, y_i^t) = M_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{6}$$

(a) (b) (c)

Figure 8. Customized mapping training samples. (a) The masks full of grids. (b) The large distortion area on the front of the sample is a triangle. (c) The trapezoidal regions can also be approximated to triangles because the influence of the marginal grids usually does not work.

In this work, the integer sampling kernel can improve the mapping performance from 2D to 3D, the relationship between input the feature map U and output map V is shown in Eq (7), where *b* rounds *x* to the nearest integer and $\delta\{\}$ is the Kronecker delta function [48], $U_{n,m}^c$ is the value at the location (*n,m*) in Channel *c* and we get a solid coordinate system of 90 cm units, as shown in Figure 9. At the same time, for the case that we need to restore the pixel information from the UV plane to the XY plane, that is, 3D to 2D, we propose a differentiable inverse transform. This sampling efficiency can improve the conversion efficiency of the camera according to the reference [49]. As shown in Eq (8), we described part of the mask fitting points in 428 keys. These key points are Delaunay-triangulated. Let the triangle set corresponding to the set $\boldsymbol{P} = \{P1,...,P2\}$ on the plane be **T**, which must meet the following requirements: (a) the endpoint of the triangle exactly belongs to the set ***P***; (b) the edges of any two triangles do not intersect or the edges of adjacent triangles coincide and (c) T forms the convex hull of ***P***. At this time, the face and mask are composed of several blocks. The more triangle blocks that are connected, the more the depth of information affects robustness. When constructing the Delaunay triangulation network and using the local optimization procedure, combine two common triangles into a polygon, and use the maximum empty circle criterion to detect whether the fourth

vertex of the polygon is within the triangle's circumscribed circle. If it is, adjust the diagonal as shown in Figure 10.

$$V_i^c = \sum_n^H \sum_m^W U_{n,m}^c \, \delta\{([x_i^s + b] - m) \times ([y_i^s + b] - n)\} \begin{cases} b = 0, x, y < 0.5 \\ b = 1, x, y > 0.5 \end{cases} \tag{7}$$

$$\frac{\partial V_i^c}{\partial U_{n,m}^c} = \sum_n^H \sum_m^W U_{n,m}^c \max(0,1 - |x_i^s - m|) \max(0,1 - |y_i^s - n|) \tag{8}$$

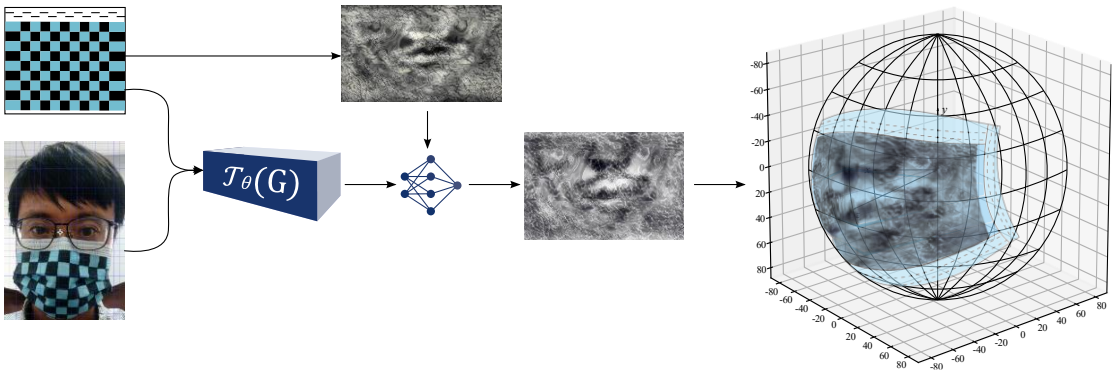


Figure 9. Mapping process from XY to XYZ and UV space. The plane template of the mask is mapped to the sample of the face with a mask through the $\mathcal{T}$ function. At the same time, the attack patch is generated according to the mask specification and a 3D attack patch appears on the mask in the 90 cm coordinate system.

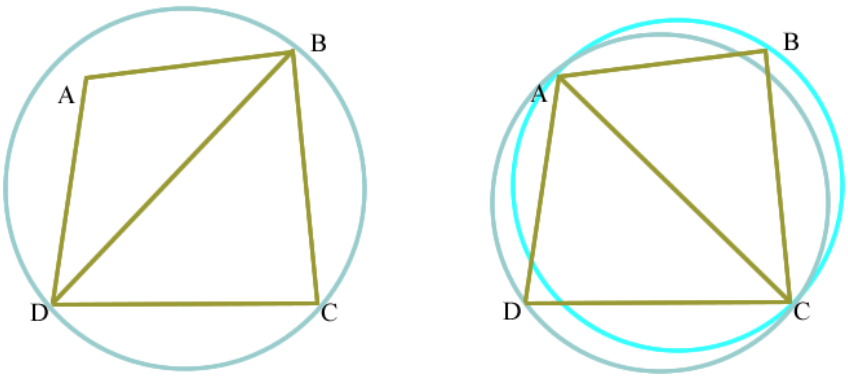


Figure 10. Local optimization procedure for Delaunay triangulation.

However, only local optimization cannot increase the accuracy of depth information. Thus, we used the interpolation method to increase the depth as shown in Figure 11: (a) Find an auxiliary window *Rect* containing the in-point set and connect the diagonal lines of *Rect* to form two triangles as the initial grid in ***P***. (b) Insert a point *P* in ***T*** and determine the triangle where *P* is located to search for adjacent triangles and circumscribe circles. (c) Delete all triangles of the circumscribed circle containing P to form a polygonal cavity containing P, and then connect the vertices of P and the cavity to form a new grid. (d) Skip (c) and repeat (b).

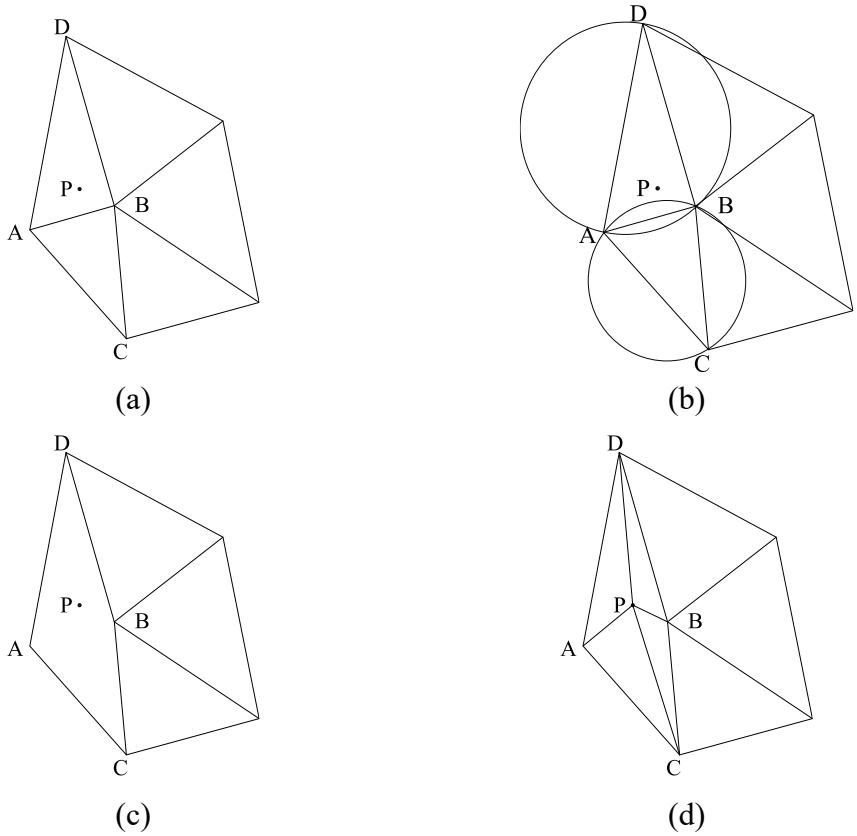


Figure 11. Point-by-point insertion algorithm for depth information. (a) Set a new node P, (b) decide how to connect P with other nodes, (c) delete the AB side and (d) Generate a new triangle block.

Target recognition needs to map the extracted 2D features into the 3D point cloud template by combining 2D calibration and RGB projection. Here, a dataset marks the point cloud coordinates, labels, colors and correlation matrix data of the wrinkle area of the PrivacyMask for a variety of faces and angles. The application of the dataset can also effectively filter the irrelevant points in two-dimensional space, prune the mapping and only keep the view cone key point cloud data. Because the angles of the camera and the face are random, the camera coordinates need to have different initial values. Therefore, the UV plane and XYZ plane need to be constantly converted to rotate from the XYZ plane to the camera's orthogonal position. First, the camera establishes a virtual XYZ coordinate system and uses the UV image to map the cone coordinate system when the UV coordinates are shifted. Then, the training results and 3D templates are used to project the UV coordinates into the XYZ coordinate system. Then, the depth information of the Z axis is rotated through the T-Net network to get the 3D coordinates of the target, as shown in Figure 12. After being verified by a multi-task convolutional neural network [3], the generated countermeasure patch can be printed on the mask, as shown in Figure 13.

The printed mask did not appear as strange or uncomfortable images, nor did it have terrible effects. It is a relative obedience pattern that makes people have no sense of conflict, although the texture on the mask looks like a catfish. The lead (Pb)-containing coating is used in the experiment. In the actual production, please use printing materials that meet local standards.

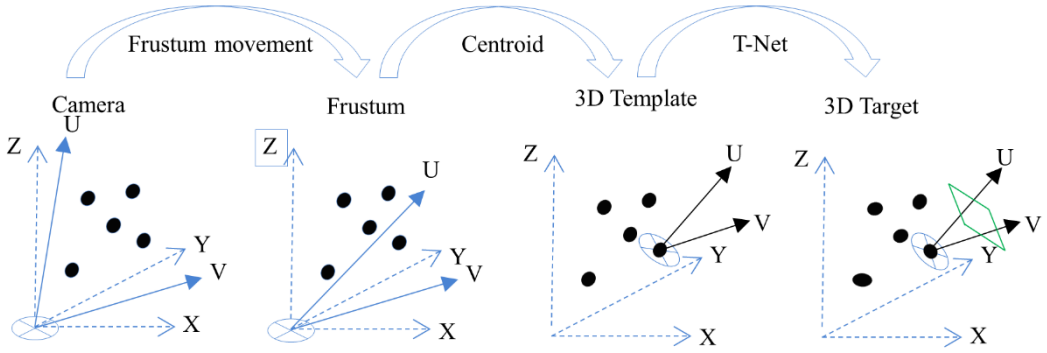


Figure 12. 2D to 3D vision conversion method.

(a) (b) (c)

Figure 13. Privacy mask. (a) Front 0 degree of the mask, (b) up 45 degrees of the mask and (c) lateral 45 degrees of the mask.

4. Experiments and results

The experiments were conducted to explore the effectiveness of algorithm fusion compatibility proof and liveness face privacy protection. We modified the Labeled Faces in the Wild (LFW) dataset and the Real World Masked Faces dataset [50] for training. The Andy_Lau dataset and masked-LFW dataset were used for testing. At the same time, 30 volunteers took part in the living test.

4.1. Attention-based approach experiment

In the experiment, we selected Baidu (BD), Arcface (AF) and Face++ (FA) as the adversary FRAs. Because these algorithms open some facial feature anchors, it is easy to add observation windows in the encoder to detect whether the generated texture is in the predetermined position.

We picked 5000 faces for the training set and registered them to the adversary FRAs. In the

first round, we fed 4000 faces in 4000 photos, and the remaining 1000 faces with over 4000 photos were trained in the second round. As an index of the training results, the false and non-match rate was evaluated; it represents the confidence in the FRA subtracted by 1. Two important findings in Figure 14 were confirmed: (1) for a single adversary FRA, the attention mechanism does not significantly increase the FNMR, but it keeps the network stable; (2) after the second training round (from 4500 minutes), the FNMR of the network with the attention mechanism increased faster.

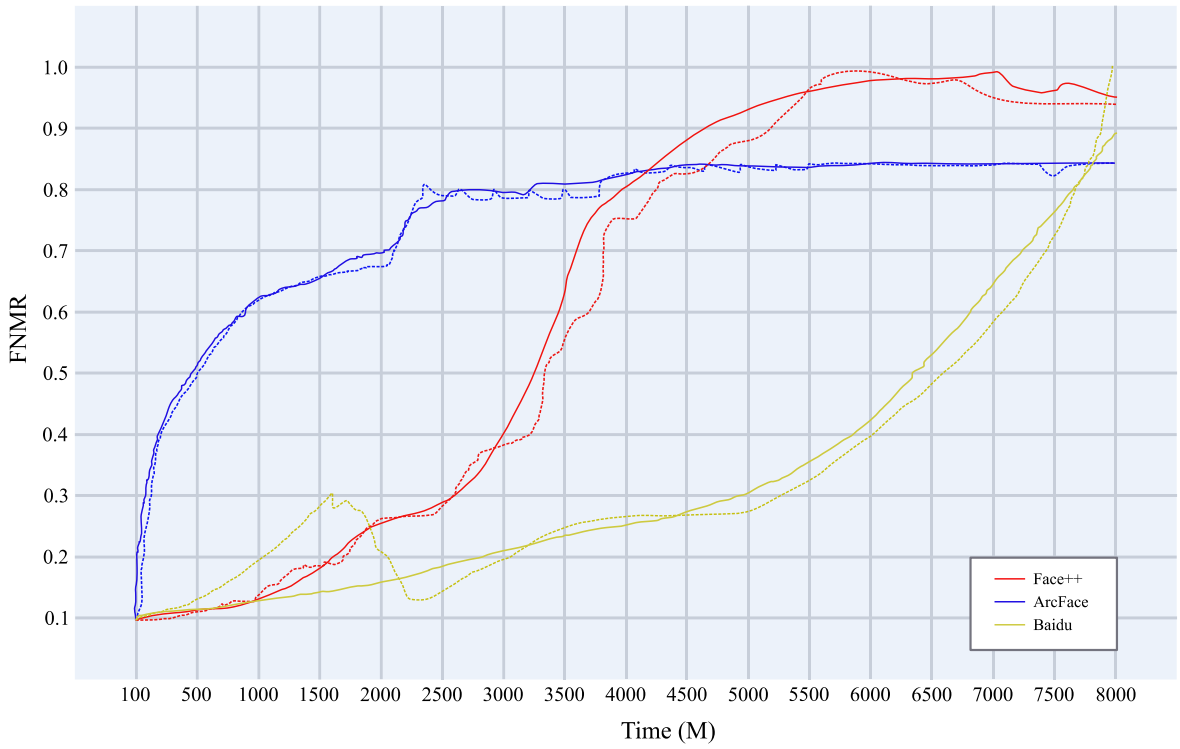


Figure 14. Application of attention mechanism in training network. Curves with different colors represent different adversary FRAs. And, the dotted line represents the generation network, and the solid line represents the network with the attention mechanism.

The experimental results show that the Baidu FRA has a big jump; at the same time, the FALSE (0.0) burst when the confidence was less than 0.05. It brings about the FNMR up to 1.0 suddenly from the 7500th minutes; clearly, this result does not accord with the principle of statistics. So, it needs to set the expected average value according to each SDK or experimental performance of FRAs. And, the following target FNMR was defined as [0.80,0.95] according to the Cannikin law [51].

We took the Baidu FRA with changeable training parameters as the baseline and connected several FRAs in parallel to the decoder. As shown in Figure 15, the training efficiency with the attention mechanism (solid line) is higher than unmodified ones (dotted line), and the training process is more rapid and stable. It makes the network more comfortable to deal with multiple FRAs. At the same time, we also confirm that it is difficult to train a good result for all closed-source FRAs using generative adversarial networks alone.

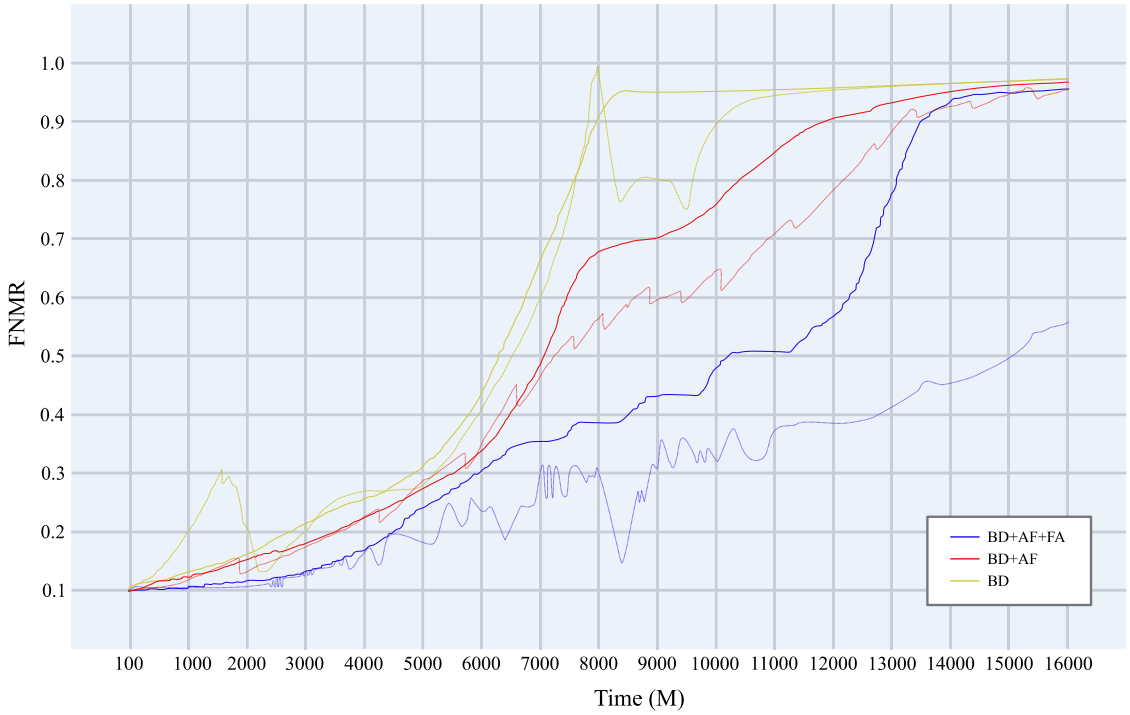


Figure 15. Attention mechanism in multi-adversary FRAs. BD stands for Baidu, AF stands for Arcface and FA symbolizes Face++. The plus sign indicates several FRAs' parallel training.

4.2. Fusion experiments

This experiment verifies whether our model has good robustness when dealing with multiple types of FRAs. Our goal is to make static face IDs invisible in Baidu, Azure face, Face++, ArcFace, Huawei, Aliyun and Tencent. The privacy protection methods compared are AdvHat (AH) [2], Pautov's eyeglasses (PE) [3], Simen's patches (SP) [4] and adversarial Patch (AP) [36].

AH uses a regular color printer to produce a rectangle sticker on the hat. An adversary sticker is created by using a unique algorithm that simulates the out-of-plane deformation of the picture of the sticker position on the hat. This approach claims that it can perplex the most sophisticated public face ID models. PE can print eyeglasses that make faces invisible in the LResNet100E-IR and ArcFace@ms1m-refine-v2 models. SP and AP cause detector classification error, so it cannot track the faces. We tested these state-of-the-art methods with the texture generated by our network to contrast the effect on the listed FRAs.

It can be inferred from Table 1 that some methods claiming to have efficient privacy protection show weaknesses in commercial FRAs. AH and PE need to train for a designative type of FRA. If they meet multiple FRAs, they cannot improve FNMRs. SP can only take effect for the target detection algorithm of a white-box model. The texture of AP cannot affect any FRA in the list. Obviously, our method is robust.

Table 1. Robustness experiments of image privacy protection methods.

Method	Framework	FNMR						
		Baidu	Azure	Face++	ArcFace	Huawei	Aliyun	Tencent
AH	LResNet100E-IR	0.038	0.015	0.013	0.213	0.011	0.012	0.017
PE	LResNet100E-IR	0.371	0.421	0.396	0.679	0.420	0.272	0.450
SP	YOLOv2	0.016	0.015	0.012	0.010	0.012	0.014	0.013
AP	EOT [52]	0.013	0.015	0.011	0.011	0.012	0.014	0.013
PM(ours)	ArcFace-v2	0.912	0.973	0.902	0.923	0.988	0.918	0.972

4.3. Liveness experiment

We invited 30 volunteers to wear masks with over 100 textures to participate in the experiment. The adversary FRAs were Baidu, Azure face, Face++, ArcFace, Huawei, Aliyun, Tencent and Apple (IOS 15.3) face ID algorithms.

An experiment on recognizing occluded faces showed that all algorithms involved have high accuracy when people wear clear masks or wear common obstructing texture decorations. These FRAs have a high recognition rate when they store a blurred picture, as shown in Figure 16. All of the FRAs can resist the pattern interference on the mask, even if the pattern is half of a face.

Except for Apple face ID, all FRAs were installed on the same hardware platform, in which the camera resolution was 2K. The volunteers conducted two groups of experiments. First, we asked each volunteer to test 500 times in an FRA and replace the PrivacyMask one at a time. Second, an FRA identified 30 volunteers in turn, and each volunteer needed to wear different PrivacyMasks each time. Finally, the average confidence was recorded, as shown in Figure 17. The experiments show that the FRAs cannot identify the subjects. People wearing the PrivacyMask can cause the FRA to mismatch face IDs. It is an excellent way to protect faces in the physical world from being tracked or authenticated. The confidence of Huawei and the Apple IOS was 0 because Huawei cannot output confidence less than 0.6 and the Apple IOS is a complete black-box model and does not output confidence.

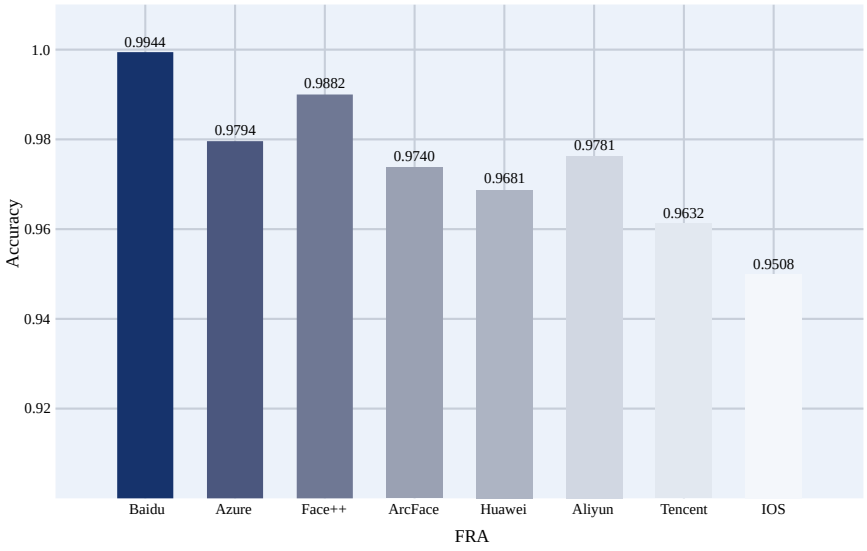


Figure 16. Accuracy experiment of face recognition without PrivacyMasks.

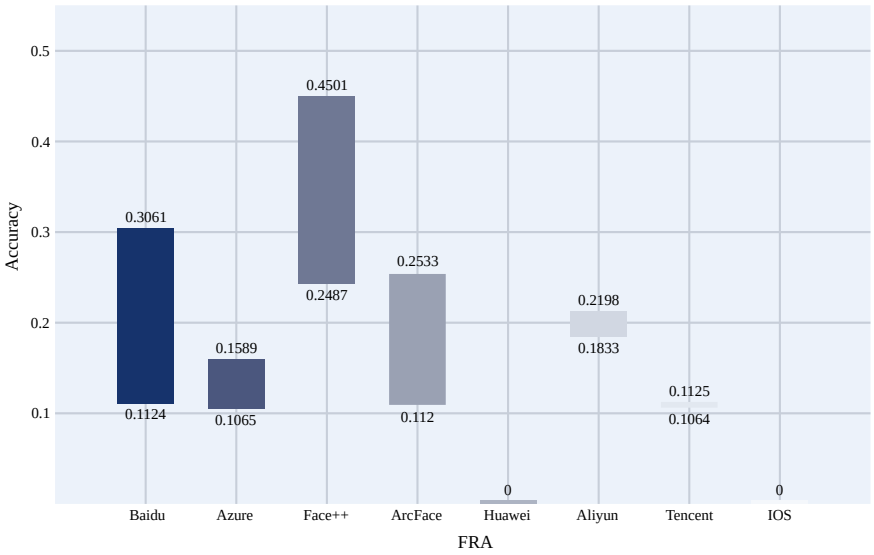


Figure 17. Accuracy experiment of face recognition with PrivacyMask.

5. Discussion

Our method is easy to commercialize and may be used by criminals to avoid monitoring equipment. Therefore, we did not release code and experimental data. The intact privacy protection contains three parts: static data privacy protection, physical world privacy protection and face recognition authorization. We have completed the first two works. The main content of the recognition authorization is that the generated facial texture carries keys and the discriminator allowed by the user will not be affected by these protective textures. After we complete the final work, there will be no way to hide face IDs illegally. Also, the public security-related camera will be able to recognize the people wearing the PrivacyMask.

6. Conclusions

PrivacyMask can generate some textures that can be printed on the mask, which makes the face detector lose its identification ability. In short, this method can protect the face in the physical world by using the attack principle. People can wear this mask to deal with face privacy snooping and face feature tracking in an unknown environment. The pattern of this mask will not put users in an awkward situation compared with the similar mask products. Our proposed texture mapping method can effectively solve the failure caused by the change of light and angle. The experimental results verified the robustness of PrivacyMask against the state-of-the-art face ID system of Baidu, Huawei and Apple, with a success rate of over 95%. Also, the attention mechanism can keep the network stable and reduce the gradient problems. It makes the idea of algorithm fusion easier to implement. Further, combined with the static data protection method we proposed before, a closed loop to protect face privacy has been formed.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, 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*, (2016), 1528–1540. https://doi.org/10.1145/2976749.2978392
2. S. Komkov, A. Petiushko, Advhat: Real-world adversarial attack on ArcFace face id system, preprint, arXiv: 1908/08705
3. M. Pautov, G. Melnikov, E. Kaziakhmedov, K. Kireev, A. Petiushko, On adversarial patches: Real-world attack on ArcFace-100 face recognition system, in *2019 International Multi-Conference on Engineering, Computer and Information Sciences*, (2019), 391–396. https://doi.org/10.1109/SIBIRCON48586.2019.8958134
4. S. Thys, W. Van Ranst, T. Goedemé, Fooling automated surveillance cameras: Adversarial patches to attack person detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, (2019), 49–55. https://doi.org/10.1109/CVPRW.2019.00012
5. N. Ud Din, K. Javed, S. Bae, J. Yi, A novel GAN-based network for unmasking of masked face, *IEEE Access*, **8** (2020), 44276–44287. https://doi.org/10.1109/ACCESS.2020.2977386
6. S. Ge, C. Li, S. Zhao, D. Zeng, Occluded face recognition in the wild by identity-diversity inpainting, *IEEE Trans. Circ. Syst. Vid.*, **30** (2020), 3387–3397. https://doi.org/10.1109/TCSVT.2020.2967754
7. R. Weng, J. Lu, Y. P. Tan, Robust point set matching for partial face recognition, *IEEE Trans. Image Process.*, **25** (2016), 1163–1176. https://doi.org/10.1109/TIP.2016.2515987
8. W. Hariri, Efficient masked face recognition method during the COVID-19 pandemic, *Signal Image Video Process.*, **16** (2022), 605–612. https://doi.org/10.1007/s11760-021-02050-w
9. D. Montero, M. Nieto, P. Leskovsky, N. Aginako, Boosting masked face recognition with multi-task ArcFace, preprint, arXiv: 2104/09874
10. M. Sharif, S. Bhagavatula, L. Bauer, M. K. Reiter, 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*, (2016), 1528–1540. https://doi.org/10.1145/2976749.2978392
11. Y. Kim, J. Na, S. Yoon, J. Yi, Masked fake face detection using radiance measurements, *J. Opt. Soc. Am. A. Opt. Image Sci. Vis.*, **26** (2009), 760–766, https://doi.org/10.1364/JOSAA.26.000760
12. N. Kose, J. L. Dugelay, Countermeasure for the protection of face recognition systems against mask attacks, in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (2013), 1–6. https://doi.org/10.1109/FG.2013.6553761
13. Y. Song, H. Zhang, A framework of face synthesis based on multilinear analysis, in *Proceedings of the 15th ACM SIGGRAPH Conference on Virtual-Reality Continuum and Its Applications in Industry*, **1** (2016), 111–114. https://doi.org/10.1145/3013971.3014026

14. J. Pu, N. Mangaokar, L. Kelly, P. Bhattacharya, K. Sundaram, M. Javed, et al., Deepfake videos in the wild: Analysis and detection, in *Proceedings of the Web Conference 2021*, (2021), 981–992. https://doi.org/10.1145/3442381.3449978
15. R. Sun, C. Huang, H. Zhu, L. Ma, Mask-aware photorealistic facial attribute manipulation, *Comput. Visual Media*, **7** (2021), 363–374. https://doi.org/10.1007/s41095-021-0219-7
16. J. Lei, Z. Liu, Z. Zou, T. Li, J. Xu, Z. Feng, et al., Facial expression recognition by expression-specific representation swapping, in *Artificial Neural Networks and Machine Learning—ICANN 2021, Lecture Notes in Computer Science*, (2021), 80–91. https://doi.org/10.1007/978-3-030-86340-1_7
17. H. Wang, G. Sun, K. Zheng, H. Li, J. Liu, Y. Bai, Privacy protection generalization with adversarial fusion, *Math. Biosci. Eng.*, **19** (2022), 7314–7336. https://doi.org/10.3934/mbe.2022345
18. W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, Z. Zhu, Robust face recognition via occlusion dictionary learning, *Pattern Recogn.*, **47** (2014), 1559–1572. https://doi.org/10.1016/j.patcog.2013.10.017
19. N. Alyuz, B. Gokberk, L. Akarun, 3-D Face recognition under occlusion using masked projection, *IEEE Trans. Inf. Foren. Sec.*, **8** (2013), 789–802. https://doi.org/10.1109/TIFS.2013.2256130
20. H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3D Face recognition under expressions, occlusions and pose variations, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2013), 2270–2283. https://doi.org/10.1109/TPAMI.2013.48
21. H. Li, D. Huang, J. M. Morvan, Y. Wang, L. Chen, Towards 3D face recognition in the real: A registration-free approach using fine-grained matching of 3D keypoint descriptors, *Int. J. Comput. Vis.*, **113** (2015), 128–142. https://doi.org/10.1007/s11263-014-0785-6
22. Y. Guo, J. Zhang, J. Cai, B. Jiang, J. Zheng, CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images, *IEEE Trans. Pattern Anal. Mach. Intell.*, **41** (2019), 1294–1307. https://doi.org/10.1109/TPAMI.2018.2837742
23. Q. Hong, Z. Wang, Z. He, N. Wang, X. Tian, T. Lu, Masked face recognition with identification association, in *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, (2020), 731–735. https://doi.org/10.1109/ICTAI50040.2020.00116
24. Y. Utomo, G. P. Kusuma, Masked face recognition: Progress, dataset and dataset generation, in *2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS)*, (2021), 419–422. https://doi.org/10.1109/ICORIS52787.2021.9649622
25. J. Prinosil, O. Maly, Detecting faces with face masks, in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, (2021) 259–262. https://doi.org/10.1109/TSP52935.2021.9522677
26. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, *Inf. Fusion*, **64** (2020), 131–148. https://doi.org/10.1016/j.inffus.2020.06.014
27. S. Avidan, M, Butman, Blind vision, in *Computer Vision–ECCV 2006*, (2006), 1–13, https://doi.org/10.1007/11744078_1
28. G. Sun, H. Wang, Image encryption and decryption technology based on Rubik's cube and dynamic password, *J. Beijing Univ. Technol.*, **47** (2021), 833–841. https://doi.org/10.11936/bjutxb2020120003
29. J. Zhou, C. Pun, Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming, *IEEE Trans. Inf. Foren. Sec.*, **16** (2021), 1088–1103. https://doi.org/10.1109/TIFS.2020.3029913

30. P. Climent-Pérez, F. Florez-Revuelta, Protection of visual privacy in videos acquired with RGB cameras for active and assisted living applications, *Multimed. Tools Appl.*, **80** (2021), 23649–23664. https://doi.org/10.1007/s11042-020-10249-1
31. K. Zheng, J. Shen, G. Sun, H. Li, Y. Li, Shielding facial physiological information in video, *Math. Biosci. Eng.*, **19** (2021), 5153–5168. https://doi.org/10.3934/mbe.2022241
32. M. Chen, X. Liao, M. Wu, PulseEdit: Editing physiological signals in facial videos for privacy protection, *IEEE Trans. Inf. Foren. Sec.*, **17** (2022), 457–471. https://doi.org/10.1109/TIFS.2022.3142993
33. S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, B. Y. Zhao, Fawkes: Protecting personal privacy against unauthorized deep learning models, in *29th USENIX Security Symposium (USENIX Security 20)*, (2020), 1589–1604
34. J. R. Padilla-López, A. A. Chaaraoui, F. Flórez-Revuelta, Visual privacy protection methods: A survey, *Expert Syst. Appl.*, **42** (2015), 4177–4195. https://doi.org/10.1016/j.eswa.2015.01.041
35. S. N. Patel, J. W. Summet, K. N. Truong, BlindSpot: Creating capture-resistant spaces, in *Protecting Privacy in Video Surveillance*, **13** (2009), 185–201. https://doi.org/10.1007/978-1-84882-301-3_11
36. T. B. Brown, D. Mané, A. Roy, M. Abadi, J. Gilmer, Adversarial patch, preprint, arXiv: 1712/09665
37. C. He, H. Hu, Image captioning with text-based visual attention, *Neural Process. Lett.*, **49** (2019), 177–185. https://doi.org/10.1007/s11063-018-9807-7
38. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2014), 1724–1734. http://dx.doi.org/10.3115/v1/D14-1179
39. Q. Guo, J. Huang, N. Xiong, MS-Pointer network: Abstractive text summary based on multi-head self-attention, *IEEE Access*, **7** (2019), 138603–138613. https://doi.org/10.1109/ACCESS.2019.2941964
40. H. Li, A. Kadav, I. Durdanovic, H. Samet, H. P. Graf, Pruning filters for efficient convnets, preprint, arXiv:1608.08710
41. J. Liu, B. Zhuang, Z. Zhuang, Y. Guo, J. Huang, J. Zhu, et al., Discrimination-aware network pruning for deep model compression, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 4035–4051. https://doi.org/10.1109/TPAMI.2021.3066410
42. H. Zou, X. Sun, 3D Face recognition based on an attention mechanism and sparse loss function, *Electronics*, **10** (2021), 2539. https://doi.org/10.3390/electronics10202539
43. J. J Koenderink, A. J. Doorn, Surface shape and curvature scales, *Image Vis. Comput.*, **10** (1992), 557–564. https://doi.org/10.1016/0262-8856(92)90076-F
44. I. G. Kang, F. C. Park, Cubic spline algorithms for orientation interpolation, *Int. J. Numer. Meth. Eng.*, **46** (1999), 45–64. https://doi.org/10.1002/(SICI)1097-0207(19990910)46:1%3C45::AID-NME662%3E3.0.CO;2-K
45. P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, et al., Overview of the face recognition grand challenge, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, **1** (2005), 947–954. https://doi.org/10.1109/CVPR.2005.268

46. M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in *Advances in Neural Information Processing Systems*, **2** (2015), 2017–2025. Available from: https://papers.nips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf
47. W. Arendt, M. Warma, Dirichlet and neumann boundary conditions: What is in between?, *J. Evol. Equations*, **3** (2003), 119–135. https://doi.org/10.1007/978-3-0348-7924-8_6
48. R. Carbó-Dorca, Logical kronecker delta deconstruction of the absolute value function and the treatment of absolute deviations, *J. Math. Chem.*, **49** (2011), 619–624. https://doi.org/10.1007/s10910-010-9781-4
49. M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, preprint, arXiv: 1506/02025
50. Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, et al., Masked face recognition dataset and application, preprint, arXiv: 2003/09093
51. X. Li, S. Liu, H. Chen, K. Wang, A potential information capacity index for link prediction of complex networks based on the cannikin, *Entropy*, **21** (2019), 863. https://doi.org/10.3390/e21090863
52. A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples, preprint, arXiv: 1707/07397

AIMS Press