Journal of Information Security and Applications 60 (2021) 102874

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

Adversarial attacks by attaching noise markers on the face against deep face recognition

Gwonsang Ryu [a], Hosung Park [b], Daeseon Choi [c,*]

[a] Department of Software Convergence, Graduate School of Soongsil University, Seoul, 07027, South Korea
[b] Department of Cyber Security and Police, Busan University of Foreign Studies, Busan, 46234, South Korea
[c] Department of Software, Soongsil University, Seoul, 07027, South Korea

ARTICLE INFO

Keywords:
Adversarial examples
Adversarial attack
Evasion attack
Face recognition
Deep neural networks

ABSTRACT

Deep neural networks (DNNs) have become increasingly effective in difficult machine learning tasks, such as image classification, speech recognition, and natural language processing. Face recognition (FR) using DNNs shows high performance and is widely used in various domains such as payment systems and immigration inspection. However, DNNs are vulnerable to adversarial examples generated by adding a small amount of noise to an original sample, resulting in misclassification by the DNNs. In this study, we attempt to deceive state-of-the-art FR by attaching noise markers on a face in the real world. To deceive an FR model in the real world, we address challenges in the attack process, including selection of locations of noise markers, the differences between colors of digital noise markers and those of noise markers after printing, the differences between the colors of noise markers that are attached to the face and those of noise markers after a picture is taken, and the differences between the locations of digital noise markers and those of noise markers that are attached to the face. In experiments, we generate noise markers considering these challenges and show that state-of-the-art FR can be deceived by attaching a maximum of 10 noise markers to a face. This can cause a security risk for FR models using DNNs.

1. Introduction

Face recognition (FR) has been a long-standing research topic in the field of computer vision and is a biometric technique that identifies facial images. FR has poor performance because facial images are collected in a real environment, and because of variation in facial expressions, light levels, distances, and image resolutions. Recently, deep neural networks (DNNs) have shown high performance in various fields, such as speech recognition [1] and natural language processing [2]. In particular, DNNs can recognize images with near-human accuracy [3]. DNNs have also been shown to exhibit good FR performance when trained with a large-scale face dataset [4–6]. Deep FR is used for various purposes, such as immigration inspection, payment services, and automated stores.

With the advancement of DNNs, there is a growing interest in their security challenges. Researchers have discovered that existing DNNs are vulnerable to adversarial attacks. Adversarial attacks generate an adversarial example by adding a small amount of noise to the original sample, resulting in misclassification by DNNs [7–10]. For example, if self-driving vehicles use DNNs, an adversarial example could cause human casualties by misrecognizing a stop sign as a speed-limit sign [11,12]. Adversarial attacks that cause misrecognition in deep FR are also being actively studied [13–16].

Representative adversarial attacks against deep FR include facial attribute attack [13], geometrically perturbed faces [14], invisible mask attack [15], and glasses attacks [16]. Facial attributes [13] generate adversarial examples in which deep FR misclassifies facial attributes as other facial attributes. For example, it can generate an adversarial example in which a deep FR misclassifies an input sample of a man as that of a woman. In geometrically perturbed faces [14], deep FR misclassifies an adversarial example generated by moving landmarks, which are facial features, as another person. Facial attribute attack [15] and geometrically perturbed faces [16] generate adversarial examples that are difficult to identify if they are modified. However, these attacks have the limitation of not being applicable in reality. An invisible mask attack [15] is an attack technique that causes deep FR to misclassify an adversary as another person by illuminating the adversary's face after attaching a light-emitting diode (LED) lens to a cap. Glasses attacks [16] are an attack technique that uses specially designed eyeglass

* Corresponding author.
E-mail addresses: gsryu@soongsil.ac.kr (G. Ryu), hspark0865@bufs.ac.kr (H. Park), sunchoi@ssu.ac.kr (D. Choi).

https://doi.org/10.1016/j.jisa.2021.102874

Available online 21 May 2021

frames to attack a deep FR. Invisible mask attack [15] and glasses attacks [16] can attack deep FR in reality; however, they need equipment or accessories to attack and cannot be applied to some services utilizing FR systems. For example, immigration inspection prohibits wearing accessories such as caps and eyeglasses when verifying identity using FR models.

In this study, we explore how to deceive deep FR models by manipulating faces without accessories, such as eyeglasses and caps. People usually think of putting on makeup as a way of manipulating their faces or making their faces appear as if they were injured. However, these methods are either noticeable or expensive, and it is difficult to paint exact colors on the face. We demonstrate technical approaches to deceive a face recognition system by attaching noise markers to a face such that they seem to be part of the original face. By attaching noise markers on the face, the adversary generates an adversarial example that only modifies the part corresponding to the locations of the markers in the digital environment to deceive deep FR models. The adversary then prints the noise markers extracted in the adversarial example and attaches the printed noise markers to his or her face. The adversary then takes a picture of his or her face and inputs it into the deep FR models so that the system misrecognizes him or her as another person. The four challenges that need to be addressed during this attack include where to place the noise markers on the face, how to minimize the differences between the colors of the digital noise markers and those of the printed noise markers, how to minimize the differences between the colors of the noise markers that are attached to the face and those of the noise markers that are taken to the camera, and how to minimize the differences between the locations of the digital noise markers and those of the noise markers attached to the face. We show that our technical approaches can deceive deep FR models by considering these challenges.

The main contributions of this study are as follows:

- We define challenges and describe technical approaches, including location selection of noise markers, color calibration, and location calibration, for deceiving deep FR models in the real world. We show that our technical approaches can deceive state-of-the-art FR models that use ring loss [17].
- We show that state-of-the-art FR models can be deceived by modifying a narrower area than in glasses attacks [16]. Glasses attacks [16] attempt to attack by modifying 6.5% of the pixels in the facial image; however, we attack by modifying at most 0.5% of the pixels in the facial image. The narrower the modified area, the harder it is to deceive the FR models.
- We perform transferability attacks to deceive black-box FR models that are trained using facial images from 107 people. To perform transferability attacks, we use noise markers generated from a white-box FR model that is trained using facial images from seven people. We show that facial images with noise markers attached can deceive black-box models.

The structure of this paper is as follows. In Section 2, we describe the background and review related works in Section 3. Technical approaches for deceiving deep FR models by attaching noise markers on the face are presented in Section 4. In Section 5, we evaluate the technical approaches. The technical approaches are discussed in Section 6. Finally, Section 7 concludes this paper.

2. Background

2.1. Deep face recognition

Recently, FR has advanced considerably because of the success of deep convolutional neural networks (CNNs), such as AlexNet [3], VGGNet [18], Google Inception Net [19], and ResNet [20]. In Deep Face [4] and DeepID [21], FR is treated as a multi-class classification problem, and deep CNN models are first introduced to train features on large multi-identity datasets. DeepID2 [22] employs identification and verification signals to achieve better feature embeddings. DeepID2+ [23] and DeepID3 [24] explored advanced network structures to boost recognition performance. FaceNet [25] uses triplet loss to learn Euclidean space embedding and a deep CNN is then trained on almost 200 million face images.

The initial Deepface [4] and DeepID [21] adopted softmax loss for feature learning. However, softmax loss usually lacks the power of discrimination. Thus, the loss function plays an important role in deep feature learning. Several loss functions have been proposed for maximizing the inter-class variance and minimizing the intra-class variance to address this problem. Contrastive loss [22] and triplet loss [25] are usually used to increase the Euclidean margin for better feature embeddings. Center loss [26] has been proposed to learn the centers for deep features of each identity, and these centers are used to reduce the intra-class variance. A large margin softmax (L-Softmax) [27] has been proposed to add angular constraints to each identity to improve feature discrimination. Angular softmax (A-Softmax) [28] improves L-Softmax [27] by normalizing the weights. CosFace [29] and ArcFace [30] introduced an additive angular cosine margin to overcome the optimization difficulty of L-Softmax and A-Softmax, respectively. Ring loss [17] encourages the norm of samples rather than explicit enforcement through a hard normalization operation. Apart from these methods, many studies have attempted to improve the performance of FR, such as VGGFace [5], Range loss [31], Normface [32], and CoCo loss [33].

2.2. Attack category against deep face recognition

Attacks against deep FR models can be divided into two categories [16]: dodging and impersonation. Both dodging and impersonation target FR models that perform multiclass classification. In particular, they attempt to find the person to whom a given face image belongs. In an impersonation attack, the adversary modifies his or her facial image to be recognized as a specific other person. The perturbed facial image is misclassified by the target FR system as a specific person chosen by the adversary. It involves a targeted attack that causes DNNs to misclassify an input corresponding to the original class as a given target class. For example, an adversary may try to disguise his or her face to be recognized as an authorized user of a laptop or smartphone that authenticates users using FR. In a dodging attack, the adversary modifies his or her facial image to be misrecognized as any other person. The perturbed facial image is misclassified by the target FR models as in any other person. It involves an untargeted attack that causes DNNs to misclassify an input corresponding to the original class as a class that is not given. For example, dodging attacks are used by benign individuals to protect their privacy against surveillance systems such as CCTVs.

In this study, we assume a white-box attack, in which an adversary has detailed information about target deep FR models. In addition, we assume that the adversary who gains access to the target deep FR models mounts an impersonation attack after the systems have been trained. That is, the adversary cannot poison the target deep FR models by altering the training data and injecting mislabeled data. Services such as immigration inspection and payment services enroll users by collecting facial images from users and verifying the users in a restricted environment over a short time. Therefore, we evaluate our technical approaches using deep FR trained from facial images collected in limited light conditions over a short time.

3. Related work

3.1. Adversarial attacks for image classification

Szegedy et al. [7] first demonstrated that the existence of small perturbations in images could fool deep DNNs into misclassification.

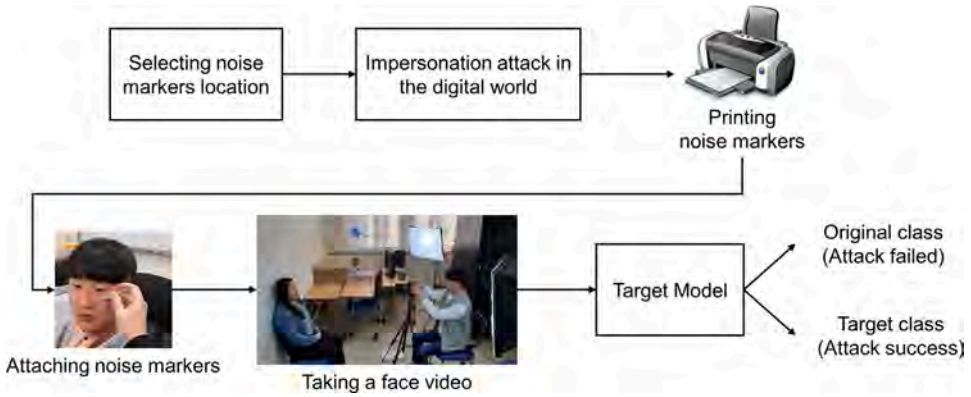

Fig. 1. The process of attacks by attaching noise markers on the face in the physical world.

They generated adversarial examples using box-constrained L-BFGS. Given an image x, L-BFGS finds a different x' that is similar to x under the Euclidean distance, but is misclassified by the DNNs. They defined a constrained minimization problem as follows:

$$\text{minimize} \quad c \cdot \|x - x'\|_2 + \text{loss}_{F,l}(x') \tag{1}$$

where $\text{loss}_{F,l}$ is a function mapping of an image to a positive real number, and it uses cross-entropy. This is done by finding the constant $c > 0$, which yields an adversarial example of the minimum distance.

Goodfellow et al. [34] proposed the fast-gradient sign method (FGSM), which can find x' through L_∞ as follows. L_∞ is the maximum pixel distance value between x and x'.

$$x' = x + \epsilon \cdot \text{sign}\left(\nabla \text{loss}_{F,t}(x)\right) \tag{2}$$

where t is a target class, $loss_{F,t}$ is the loss function of the target DNNs, and ϵ is a small constant value that restricts the norm of the perturbation.

Kurakin et al. [35] introduced iterative FGSM (I-FGSM), which is an extension of FGSM. Instead of updating the amount ϵ in every step, a smaller amount, α, was changed and it was eventually clipped by the same ϵ, as follows.

$$x'_i = x'_{i-1} - \text{clip}_\epsilon\left(\alpha \cdot \text{sign}\left(\nabla \text{loss}_{F,t}\left(x'_{i-1}\right)\right)\right) \tag{3}$$

As I-FGSM generates a fine-tuned adversarial example during a given iteration on a particular model, it has a higher attack success rate as a white box attack than FGSM.

Papernot et al. [8] introduced an attack optimized under the L_0 distance, which is known as the Jacobian-based saliency map attack. This is a simple iterative method for a targeted attack. It finds a component that reduces the adversarial example's saliency value to induce the minimum distortion. The saliency value is a measure of the importance of an element in determining the output class of a model.

Moosavi-Dezfooli et al. [9] proposed Deepfool, which generates an adversarial example more efficiently than FGSM; the example is similar to the original image. The method looks for x' using the linearization approximation method on a DNN to generate an adversarial example. However, because DNNs are not completely linear, and the method requires multiple iterations, Deepfool is a more complicated process than FGSM.

Carlini and Wagner [10] introduced a set of adversarial attacks that make the perturbations quasi-imperceptible by restricting L_0, L_2, and L_∞ norms. The L_2 attack of these attacks optimizes an objective function as follows:

$$\text{minimize} \quad \|x - x'\|_2 + c \cdot f_Z(x') \tag{4}$$

where f_Z is defined as follows:

$$f_Z(x') = \max\left(\max\left(Z\left\{x'_i : i \neq t\right\}\right) - Z(x')_t, 0\right) \tag{5}$$

where Z is a logits vector that is the output of all layers except the softmax activation function and t is a target class. Carlini and Wagner [10] suggested a method to control the attack success rate, even with some increased distortion, by incorporating a confidence value.

3.2. Adversarial attacks against deep face recognition

Rozsa et al. [13] proposed an attack technique for DNNs with a so-called fast flipping attribute. They found that the robustness of DNNs against adversarial attacks varies highly between facial attributes. It is claimed that adversarial attacks are effective in changing the label of a target attribute to a correlated attribute. Mirjalili and Ross [36] proposed a technique that modifies a facial image such that its gender is modified. However, its biometric utility for an FR system remains intact. Shen et al. [37] proposed two different techniques to generate adversarial examples for faces that can have high 'attractiveness scores' but low 'subjective scores' for face attractiveness evaluation using DNNs. Daboueo et al. [14] proposed an attack technique in which a deep FR system misclassifies a facial image generated by moving landmarks that are facial features.

Zhou et al. [15] proposed an attack technique called the invisible mask. Invisible mask attacks attack deep FR models such that they misclassify an adversary as another person by illuminating the adversary's face by infrared light after attaching an LED lens. This type of attack considers the shape, color, size, location, and brightness of infrared light. Sharif et al. [16] attacked deep FR models by printing adversarial perturbations on the frames of eyeglasses. An adversary will be misclassified as a target person chosen by the adversary when the adversary wears certain eyeglasses.

4. Technical approach

The process of deceiving deep FR models by attaching noise markers on the face is as shown in Fig. 1. An adversary first selects the location of noise markers for modification and then generates an adversarial example by only modifying pixel values corresponding to the locations of the noise markers. The adversary prints the noise markers and then attaches them to the locations of his or her face corresponding to the selected locations. The adversary takes a photo of his or her face with noise markers attached and inputs it into the FR models to check whether the attack succeeded or failed. In this section, we define the three challenges that need to be solved to deceive the deep FR models and address technical approaches to solve the defined challenges.

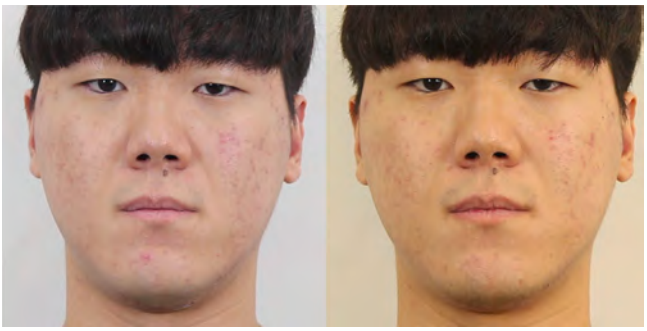

Fig. 2. Example of color differences caused by changes in brightness of sunlight.

4.1. Challenge definition

First, we generate an adversarial example of a successful attack by placing noise markers on the image of the adversary's face in the digital environment. In this process, the question of where to place markers remains. The adversary cannot place the markers on image parts other than the face, including the background and the hair, because the adversary has to attack deep FR models by attaching noise markers on his or her face. In addition, the adversary does not know how many noise markers he or she has to generate to deceive the FR models.

Second, facial images always change even when the image of the face is taken in the same environment because of changes in the brightness due to environmental changes, such as weather and time of day, as shown in Fig. 2. In addition, the adversary cannot attach noise markers to precisely the same locations as those of noise markers of an adversarial example when the adversary attaches the noise markers on his or her face.

Third, the adversary tries to attack by attaching noise markers on the face after printing the noise markers. However, when the adversary prints digital noise markers, the colors of the printed noise markers are not exactly the same as the colors of the digital noise markers. In addition, when digital noise markers are printed and photographed by attaching printed noise markers to the face, there is a difference between the colors of successful digital noise markers and the colors of the captured noise markers. Therefore, the adversary has to generate noise markers that minimize the color differences to deceive FR models by attaching markers on the face.

4.2. Selecting noise marker location

We consider three methods to select the location of noise markers on the face: random selection, location selection that significantly affects the FR system, and the location selection with the most noise in the adversarial example, where only the facial area is modified. The locations of noise markers in the face are given as follows.

$$a_i \in \{a_0, a_1, \ldots, a_{n-1}\} \tag{6}$$

where n is the number of possible locations of noise markers in the facial area, and a_i is a vector that only has values of 0 and 1. The vector has a value of 1 if the part corresponds to the location of the noise marker and 0 for the rest of the locations. In addition, noise markers do not overlap with the other noise markers.

First, the location of the noise marker that can be taken in the facial area is determined by random selection. An integer is extracted by a uniform distribution, and the location of the noise marker is then selected by adding the row and column size of the noise marker to the extracted integer. An adversarial example x'_{a_i} is then generated by modifying only the pixel values corresponding to the selected location of the adversary's facial sample as follows.

$$x'_{a_i} = x + \delta \cdot a_i \tag{7}$$

where x is the adversary's facial sample and δ represents a noise vector. An adversarial example that only modifies the pixel values corresponding to the location of the noise markers is generated by minimizing the objective function as follows:

$$f_Z\left(x'_{a_i}\right) = \max\left(\max\left(Z\left\{x'_{a_{i_k}} : k \neq t\right\}\right) - Z\left(x'_{a_i}\right)_t, 0\right) \tag{8}$$

where Z is a logits vector that is the output of all layers except the softmax activation function, and t is a target label.

Second, the location selection that affects the FR system the most generates adversarial examples for the location of all markers that can be placed in the facial area in the digital environment. An adversary then calculates the differences between the logits vector of the original attack sample and that of adversarial examples when inputting the FR system as follows.

$$\underset{i}{\text{argmax}} \quad \left\| Z(x) - Z\left(x'_{a_i}\right) \right\|_1 \tag{9}$$

where x is an original attack sample and $\|\cdot\|_1$ denotes the L_1 norm. The location of the noise marker is selected as the location with the largest difference from the logits vector of the original attack sample.

Third, the adversary generates an adversarial example where pixel values corresponding to the facial area are modified. The adversarial example is generated by using an L_2 attack [10]. The adversary then calculates the amount of noise for all the locations of the noise markers, and the location with the most noise is selected as follows.

$$\underset{i}{\text{argmax}} \quad \|\delta \cdot a_i\|_1 \tag{10}$$

where δ is the noise vector of an adversarial example where only the pixel values corresponding to the facial area are modified. Fig. 3 shows the location and order of noise markers selected by each method.

4.3. Color and location calibration

To minimize the color differences caused by the change in brightness due to environmental changes, such as weather and time of day, we generate adversarial examples by adding the variable for the color margin as follows.

$$x'_{\beta_0} = x'_{a_i} - \beta \cdot a_i, \qquad x'_{\beta_1} = x'_{a_i} + \beta \cdot a_i \tag{11}$$

where β is a variable for the color margin. We then look for a single perturbation such that the target deep FR system misclassifies x'_{a_i}, x'_{β_0}, and x'_{β_1} as the target t by minimizing the objective function as follows.

$$\text{minimize} \quad f_Z\left(x'_{a_i}\right) + f_Z\left(x'_{\beta_0}\right) + f_Z\left(x'_{\beta_1}\right) \tag{12}$$

A single perturbation generated by Eq. (12) may cause attacks to succeed even with a slight color difference. The single perturbation includes several noise markers. We call this technique color calibration.

To minimize the differences between the locations of the noise markers of the digital adversarial example and the attached noise markers on the face, we look for a single perturbation using multiple attack samples such that the target deep FR system misclassifies a set of adversarial examples as target t as follows.

$$\underset{i}{\text{argmin}} \quad \sum_{x \in X} f_Z\left(x + \delta \cdot a_i\right) \tag{13}$$

where X is a set of an adversary's facial samples. The target FR system misclassifies all adversarial examples generated as the target t by adding a single perturbation. For example, if the adversary finds a single perturbation for the five attack samples, five adversarial examples with the same perturbation are generated. The single perturbation may minimize the location differences because a set of attack samples involves a slight location difference. We refer to this technique as

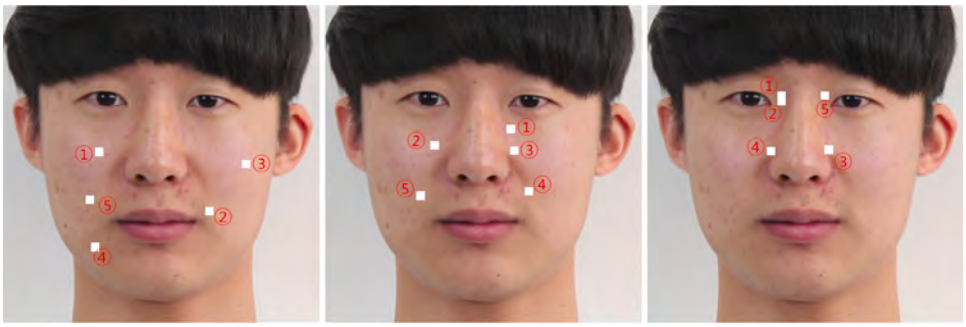

Fig. 3. Example of location and order of noise markers selected by (a) random selection, (b) location selection that significantly affects the FR system, and (c) location selection with the highest amount of noise in the adversarial example.

location calibration. We look for a single perturbation by minimizing the objective function to apply the color and location calibration concurrently, as follows.

$$\underset{i}{\text{argmin}} \sum_{x \in X} \left(f_Z\left(x'_{a_i}\right) + f_Z\left(x'_{\beta_0}\right) + f_Z\left(x'_{\beta_1}\right) \right) \tag{14}$$

To minimize the differences between the colors of digital noise markers and noise markers taken after printing, we create a color mapping table that maps the digital color table with 5,832 colors, and we create a color table taken after printing, as shown in Fig. 4. Let the digital color table be P, the color table taken after printing be $\hat{P}$, and a single pixel of noise markers be $\hat{c}$. P and $\hat{P}$ consist of 5,832 RGB triplets with values between 0 and 1, and $\hat{P}$ also has values between 0 and 1. To minimize the differences between the colors of the digital noise markers and noise markers taken after printing, we select $\hat{p} \in \hat{P}$ with $|\hat{c} - \hat{p}|$ and then print $p \in P$ such that $|\hat{p} - p|$ is minimized. This process is performed for all pixels of all noise markers. Fig. 5 shows the color differences of noise markers printed before and after using the color mapping tables.

5. Evaluation

5.1. Experiment setup

Seven people from our laboratory volunteered to collect data to evaluate our technical approaches. The dataset contained two women and five men aged 22 to 37 years. We collected face datasets by capturing videos to collect natural faces, including facial changes and facial movements. We used a Canon EOS 650D camera to shoot the facial video at a resolution of 1920 × 1080, and the facial video was shot at 25 fps. To minimize color changes due to brightness changes of light according to external environmental changes, we installed two lights, as shown in Fig. 1. Subjects sat at a fixed distance from the camera and shot facial videos.

We performed two sessions for all subjects to collect facial videos. In the first session, we shot face videos without allowing subjects to blink their eyes, move their face, or change their facial expressions. Unlike the first session, in the second session, we shot facial videos while allowing subjects to blink their eyes, move their faces, and change their facial expressions. Facial videos of each subject were taken for approximately 14 s per shot and were collected two times daily for five days in two sessions. For all frames of all facial videos, we cropped and aligned faces using multi-task CNNs [38] and resized the facial samples to 224 × 224 pixels.

We trained three FR models to evaluate their performance according to the number of days of data collection and our technical approaches. DNN_A was trained using eight facial videos, each collected on the first and second days in the two sessions, and evaluated four facial videos collected on the fifth day in the two sessions. We also trained DNN_B using facial videos collected on the second and fourth days in the same way as DNN_A. The facial videos were collected two times each for both sessions on four days; these videos were used to train DNN_C, and the facial videos that were collected two times each for both sessions on the fifth day were used to evaluate DNN_C. Therefore, DNN_C was trained using 16 facial videos for both sessions, and we evaluated the DNN_C using four facial videos for both sessions. DNN_A, DNN_B, and DNN_C were trained using Ring loss [17] as the loss function and ResNetV2 50 layers [39]. Both DNN_A and DNN_B showed 99.24% and 99.16% accuracy for the test dataset, and DNN_C showed 99.43% accuracy for the test dataset.

Since DNN_A, DNN_B, and DNN_C are trained using a small dataset, we trained two additional FR models using all facial images collected from seven people and a K-Face sample dataset [40], which contains 144,000 facial images from 100 Koreans and contains 55 women and 45 men aged 20 to 60. We split all facial images into a training dataset and a test dataset at a ratio of 7:3. Then, we trained DNN_D using MobileNetV2 [41] and DNN_E using Inception-ResNetV2 [42]. DNN_D showed 98.99% accuracy and DNN_E showed 99.0% accuracy for the test dataset. We attempted transferability attacks against DNN_D and DNN_E using cases of successful attacks against DNN_A, DNN_B, and DNN_C in the real world. We used the Tensorflow library for Python to train DNN_A, DNN_B, DNN_C, DNN_D, and DNN_E, and to test our attacks.

5.2. Selection of noise marker location

To evaluate the three methods for selecting the location of noise markers, we tried to attack by adding markers one at a time until the attack was successful. Fig. 6 shows the process by which the adversary generates an adversarial example with a limited number of noise markers in the digital environment. We attempted to attack with 3 × 3 pixel and 4 × 4 pixel noise markers; however, because the attacks were not successful, we set the size of the noise markers to 5 × 5 pixels. We set the maximum number of noise markers to 50. If the attack was not successful with 50 noise markers, it was considered a failure. We used DNN_C to evaluate three methods for selecting the location of noise markers. The number of attack samples used per person was 10. The 10 attack samples were extracted from facial videos taken without allowing subjects to blink their eyes, move their face, or change their facial expressions. We attempted impersonation attacks that cause the target FR model to misrecognize a subject as one of the other six subjects. For each subject, we attempted an impersonation attack in which the target FR model misclassified one subject as one of the other six subjects. The attacks were attempted 60 times per subject, and the attacks were attempted 420 times for all subjects.

Random selection. We randomly selected the locations of the noise markers using a uniform distribution. The randomly selected location each noise marker was extracted again if the location was not in the facial area or overlapped with the other extracted noise markers. As a

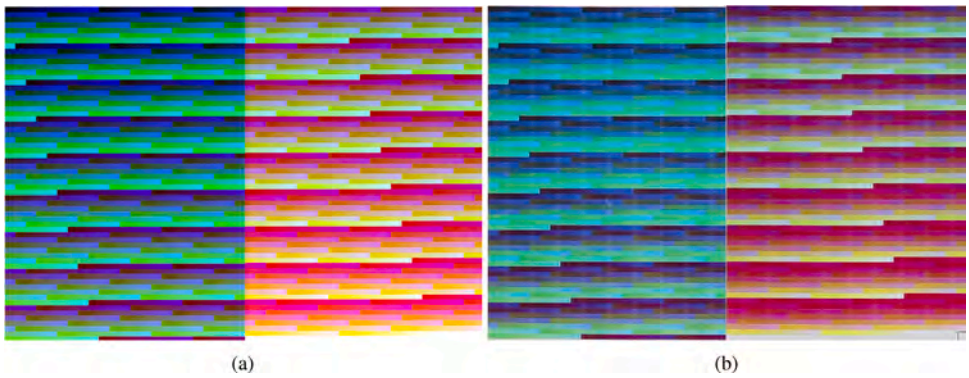

Fig. 4. The color mapping tables: (a) a color table on the digital environment, (b) a color table taken after printing. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

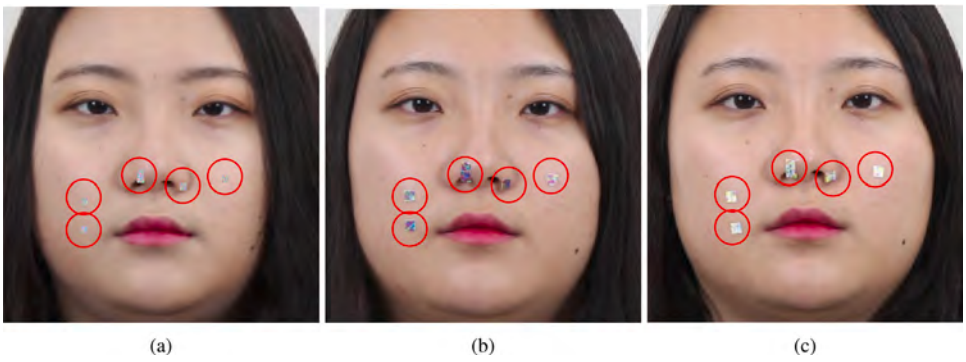

Fig. 5. The comparison of color differences of noise markers printed before and after using the color mapping tables: (a) digital noise markers, (b) noise markers attached to the face before using the color mapping tables, (c) noise markers attached to the face after using the color mapping tables. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

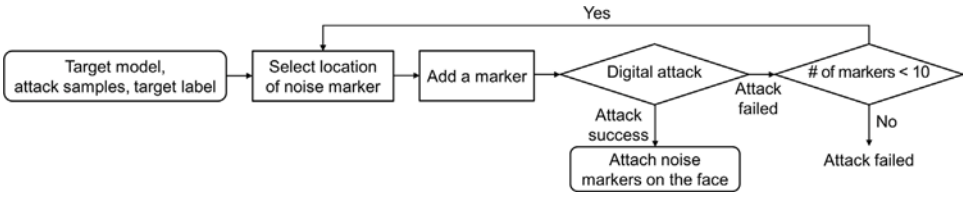

Fig. 6. The attack process to deceive deep FR models by attaching noise markers on the face in the digital environment.

Table 1
Experimental results of basic digital attacks.

Target model	# of attacks	# of successful attacks	Attack success rate
DNN_A	42	6	14.29%
DNN_B		2	4.76%
DNN_C		5	11.9%

Table 2
Experimental results of the digital attacks applying color calibration.

Target model	# of attacks	# of successful attacks	Attack success rate
DNN_A	42	3	7.14%
DNN_B		1	2.38%
DNN_C		4	9.52%

result of experiments using the random selection method, an average of 41.2 noise markers for the successful attack were required. The reason why random selection required a lot of noise markers for successful attacks is that even if a selected location is modified, the FR models are not significantly affected. Therefore, random selection is not an effective method.

Location selection most affecting the FR model. We attempted the attack by placing noise markers one by one for every facial area of the attack samples. We then attacked by modifying the location where the change in DNN_C's logits vector was the largest. As a result, an average of 24.6 noise markers were required for a successful attack. Location selection most affecting the FR model was more effective than random selection. However, it takes a long time because it is necessary to generate adversarial examples by placing noise markers one by one for every facial area of the attack samples.

Location selection with the most noise in adversarial example. We first generated an adversarial example that only modified pixel values corresponding to the facial area of the original attack sample. The amount of noise was then calculated by considering the marker size for all facial areas of the generated adversarial example. We then attacked by modifying the attack sample, adding markers with the most noise one at a time. As a result, an average of 22.3 noise markers were found to be required for a successful attack. Location selection with the most noise in the adversarial example is more effective than

Table 3
Experimental results of the digital attacks applying location calibration.

# of attack samples	Target model	# of attacks	# of successful attacks	Attack success rate
5	DNN_A	42	3	7.14%
	DNN_B		0	0%
	DNN_C		2	4.76%
10	DNN_A		3	7.14%
	DNN_B		1	2.38%
	DNN_C		2	4.76%

Table 4
The experimental results of the digital attacks applying color and location calibration.

# of attack samples	Target model	# of attacks	# of successful attacks	Attack success rate
5	DNN_A	42	2	4.76%
	DNN_B		0	0
	DNN_C		1	2.38%
10	DNN_A		3	7.14%
	DNN_B		0	0%
	DNN_C		1	2.38%

random selection and location selection with the greatest effect on the FR model. In addition, this approach takes less time than location selection with the greatest effect on the FR model because it only generates one adversarial example and then calculates the amount of noise considering the size of the noise markers for all facial areas. Thus, we performed the following experiments using location selection methods with the most noise.

5.3. Experiments in digital environment

Our attack method is based on an adversarial example of a successful digital attack, and the attack must be performed by attaching noise markers on the face. Thus, we limited the maximum number of markers to 10 because others could notice if the number of markers increased. For each subject, we attempted an impersonation attack in which the target FR model misclassified one subject as one of the other six subjects. The attacks were attempted 42 times for all subjects.

Basic attack. As a result of experimenting using basic digital attacks, we observed a 14.29% attack success rate against DNN_A, a 4.76% attack success rate against DNN_B, and an 11.9% attack success rate against DNN_C. The attack success rates were low because the FR model was deceived by modulating the location of a maximum of 10 markers. Table 1 presents the experimental results for basic digital attacks.

Color calibration. We set the variable for the color margin to 3/255. We choose this setting because it was the maximum value that we could set, as the attack success rate decreased as the color margin increased. As a result of experimental digital attacks using color calibration, a 7.14% attack success rate against DNN_A, a 2.38% attack success rate against DNN_B, and a 9.52% attack success rate against DNN_C were observed. We showed that the digital attacks applying color calibration had lower success rates than basic digital attacks. Table 2 represents the experimental results for digital attacks applying color calibration.

Location calibration. We searched for a single perturbation using 5 and 10 attack samples such that the target deep FR model misclassifies a set of adversarial examples as a target t. As a result of experimentation with digital attacks using location calibration, we found a 7.14% attack success rate against DNN_A, a 0% attack success rate against DNN_B, and a 4.76% attack success rate against DNN_C when using five attack samples. In addition, a 7.14% attack success rate against DNN_A, a 2.38% attack success rate against DNN_B, and a 4.76% attack success rate against DNN_C were observed when using 10 attack samples. Table 3 presents the experimental results for digital attacks applying location calibration.

Color and location calibration. We attempted a digital attack that applied color and location calibration. For the digital attack using 5 attack samples and the color margin, 15 adversarial examples were generated by a single perturbation. As a result of experimentation using a digital attack with 5 attack samples and the color margin, we found a 4.76% attack success rate against DNN_A, a 0% attack success rate against DNN_B, and a 2.38% attack success rate against DNN_C. As a result of experimenting using a digital attack with 10 attack samples and the color margin, we found a 7.14% attack success rate against DNN_A, a 0% attack success rate against DNN_B, and a 2.38% attack success rate against DNN_C. Table 4 presents the experimental results for digital attacks applying color and location calibration.

5.4. Experiments in the real world

We printed noise markers using a color mapping table for an adversarial example of a successful digital attack and attached it to the face to deceive the target FR model. We then recorded facial videos, allowing eye blinks, facial movements, and facial expression changes in the second session. For all frames of the facial videos taken, we cropped and aligned faces and resized them to 224 × 224, which was similar to the dimensions of the facial video sets that were used for training and testing the DNNs. We then evaluated the attack performance for all facial samples.

Basic attack. As a result of experimenting using attacks by attaching noise markers on the face for adversarial examples of successful basic digital attacks, the three FR models correctly recognized the faces in most cases, and the attacks were successful in some cases. In cases of attacks attempting to misclassify $Subject_3$ as $Subject_4$, $Subject_4$ as $Subject_6$, and $Subject_5$ as $Subject_6$ against DNN_A, DNN_A misclassified the frames. In addition, in cases of attacks presenting $Subject_3$ as $Subject_5$ against DNN_A and attacks presenting $Subject_0$ as $Subject_2$ against DNN_C, we showed that there were many frames that the FR models misclassified as someone other than the target subject, which was unintentional. This is because the objective function minimizes the difference between logit values corresponding to the original and target classes, while also reducing the differences between the logit values corresponding to other classes. $Subjects$ represent those who participated in the experiment.

Color calibration. In attack experiments using noise markers attached to the face in adversarial examples with successful digital attacks applying color calibration, we showed that the three FR models misclassified more frames than they did for basic attacks, which proves that the noise markers generated by applying the variable for the color margin can minimize the slight color changes. Table 5 represents attack results before and after applying color calibration.

Table 5
Comparison of attack results before and after applying color calibration.

Target model	Case of attack success	Basic attack			Color calibration		
		org	target	others	org	target	others
DNN_A	$Subj_3 \rightarrow Subj_4$	243	0	14	0	0	227
	$Subj_3 \rightarrow Subj_5$	0	0	266	0	0	286
	$Subj_4 \rightarrow Subj_6$	264	1	0	0	284	0
DNN_B	$Subj_4 \rightarrow Subj_6$	62	213	0	11	261	0
DNN_C	$Subj_0 \rightarrow Subj_1$	199	118	0	187	139	0
	$Subj_5 \rightarrow Subj_1$	321	0	0	0	0	278
	$Subj_5 \rightarrow Subj_2$	319	0	0	0	285	0
	$Subj_5 \rightarrow Subj_3$	324	0	0	0	0	285

Location calibration. In attack experiments using noise markers attached to the face in adversarial examples with successful digital attacks applying location calibration, the noise markers generated using 5 attack samples and 10 attack samples showed similar results against DNN_A; however, they did not show similar results against DNN_C. In cases of attacks attempting to misclassify $Subject_3$ as $Subject_4$, $Subject_3$ as $Subject_5$, and $Subject_4$ as $Subject_6$ against DNN_A, DNN_A misclassified all frames as other subjects rather than the original subject when using 5 and 10 attack samples. In cases of attacks attempting to misclassify $Subject_0$ as $Subject_1$ against DNN_C, DNN_C misclassified some frames as the target subject when using 5 attack samples, but DNN_C misclassified many frames as the target subject when using 10 attack samples. In contrast, in the case of attacks presenting $Subject_5$ as $Subject_1$ against DNN_C, DNN_C misclassified many frames as the other subjects rather than the target subject when using 5 attack samples. However, DNN_C only misclassified one frame as the target subject when using 10 attack samples. Table 6 represents attack results before and after applying location calibration.

Color and location calibration. In attack experiments using noise markers attached to the face in adversarial examples with successful digital attacks applying color and location calibration, in all cases of attacks against DNN_A, DNN_A misclassified all frames as other subjects. In addition, in cases of attacks attempting to misclassify $Subject_0$ as $Subject_1$ against DNN_C, we showed that DNN_C increased the number of frames misclassified as the target when the number of attack samples was 10 rather than 5. Table 7 represents the attack results before and after applying color and location calibration. The sections marked with '–' from Tables 5–7 indicate cases where the digital attack failed. Fig. 7 shows the attack results for physical attacks applying color and location calibration. The target deep FR models recognized original samples without adding noise markers as the original subject. However, the target deep FR models misrecognized facial samples when attaching noise markers as the target subject.

5.5. Transferability attacks

In order to perform transferability attacks, we used cases of successful attacks by applying color and location calibration in the real world. The cases include attacks attempting to misclassify $Subject_3$ as $Subject_4$, $Subject_3$ as $Subject_5$, and $Subject_4$ as $Subject_6$ against DNN_A, and $Subject_0$ as $Subject_1$ against DNN_C. We attempted transferability attacks by inputting adversarial examples of each case into DNN_D and DNN_E. Transferability attacks based on basic attacks rarely deceived both DNN_D and DNN_E. In all cases of transferability attacks, the number of adversarial frames misclassified as the target or other subjects decreased compared to the results of attacking each target model. The reason that the number of misclassified adversarial frames decreased is that DNN_D and DNN_E were trained using a larger dataset and DNN architectures other than those of DNN_A and DNN_C. We were still able to deceive DNN_D and DNN_E using adversarial frames by applying color and location calibration. Table 8 represents results of transferability attacks against DNN_D and Table 9 represents results of transferability attacks against DNN_E.

6. Discussion

In this study, we addressed challenges such as the color and location differences of noise markers that affect the process of attacking deep FR models by attaching noise markers to faces in the real world. We showed that deep FR models could be deceived by attaching small noise markers on the face. In this section, we discuss the advantages and disadvantages of our technical approaches by comparing them to glasses attacks [16].

There are two advantages to our method. First, we succeeded in deceiving deep FR models by attaching a maximum of 10 noise markers of 5×5 pixels on the face in the real world. Because we modulated a maximum of 0.5% of the pixels of the 224×224 pixel facial images, we perturbed a narrower area than that of glasses attacks [16], which perturb approximately 6.5% of the pixels of the 224×224 pixel facial images. In terms of the number of pixels, 6.5% of a 224×224 image is approximately 3,261 pixels, and the maximum 10 noise markers of 5×5 pixels correspond to a maximum of 250 pixels. Adversarial attacks are challenging if the adversary modifies a few pixels. Fig. 8 shows adversarial facial images generated by a glasses attack and by our attack. Second, we created a color mapping table by combining 5,832 colors to represent more sophisticated colors. In addition, we used a variable for the color margin and generated a single robust perturbation using multiple attack samples to minimize the slight color and location differences of noise markers. A glasses attack [16] applies a non-printability score (NPS) to generate perturbations in color combinations that can be printed to minimize the differences between the digital color and printed color of eyeglass frames. In addition, it minimizes color differences using a color mapping table including 30 colors. However, the NPS is difficult to apply because the colors that can be expressed in different printers are difficult to identify accurately, and in the color mapping table, it is challenging to minimize the differences between digital colors and printed colors. Therefore, we are able to represent more sophisticated colors than NPS, and we confirmed that our technical approaches were effective.

There are two disadvantages to our method. First, our experimental results showed a low attack success rates from digital attacks. To increase the attack success rate, we can increase the number of noise markers. The attack success rate when attacking the FR model using the maximum 50 markers was 90.32%. However, we limited the number of noise markers to 10 because attaching 50 noise markers is very noticeable. Despite limiting the number of noise markers to 10, when noise markers are attached to the face, they are noticeable. Therefore, another method is needed to deceive the FR model without being noticed. Second, we performed experiments to attack a deep FR model using facial videos collected in a restricted space. This approach is only applicable for FR models deployed within buildings. However, other practical scenarios are more challenging, and effective attacks may have to be tolerant to a wider range of imaging conditions.

7. Conclusion

In this study, we defined the challenges that must be solved when deceiving deep FR models by attaching noise markers on faces in the real world and demonstrated technical approaches to minimize color differences and location differences of the noise markers. In addition, we evaluated three methods for selecting the locations of the noise markers and showed that the deep FR system could be deceived by attaching a maximum of 10 noise markers of 5×5 pixels on the face in the real world. In future research, we will recruit more participants and conduct the study with face datasets collected under various shooting environments using different devices. In addition, we will attempt to

Table 6
Comparison of attack results before and after applying location calibration.

Target model	Case of attack success	Basic attack			5 attack samples			10 attack samples		
		org	target	others	org	target	others	org	target	others
DNN_A	$Subj_3 \rightarrow Subj_4$	243	0	14	0	0	316	0	0	268
	$Subj_3 \rightarrow Subj_5$	0	0	266	0	0	275	0	0	314
	$Subj_4 \rightarrow Subj_6$	264	1	0	0	275	0	0	272	0
DNN_B	$Subj_3 \rightarrow Subj_1$	228	44	0	–	–	–	0	272	0
DNN_C	$Subj_0 \rightarrow Subj_1$	199	118	0	303	20	0	221	102	0
	$Subj_5 \rightarrow Subj_1$	321	0	0	11	0	254	319	1	0

Table 7
Comparison of attack results before and after applying color and location calibration.

Target model	Case of attack success	Basic attack			Color & 5 attack samples			Color & 10 attack samples		
		org	target	others	org	target	others	org	target	others
DNN_A	$Subj_3 \rightarrow Subj_4$	243	0	14	0	0	278	0	0	282
	$Subj_3 \rightarrow Subj_5$	0	0	266	–	–	–	0	0	285
	$Subj_4 \rightarrow Subj_6$	264	1	0	0	286	0	0	284	0
DNN_C	$Subj_0 \rightarrow Subj_1$	199	118	0	173	157	0	116	221	0

Table 8
Results of transferability attacks against DNN_D before and after applying color and location calibration.

Target model	Case of attack success	Basic attack			Color & 5 attack samples			Color & 10 attack samples		
		org	target	others	org	target	others	org	target	others
DNN_A	$Subj_3 \rightarrow Subj_4$	257	0	0	269	0	9	258	0	24
	$Subj_3 \rightarrow Subj_5$	262	0	4	–	–	–	270	0	15
	$Subj_4 \rightarrow Subj_6$	265	0	0	276	10	0	253	15	0
DNN_C	$Subj_0 \rightarrow Subj_1$	310	7	0	316	14	0	295	42	0

Table 9
Results of transferability attacks against DNN_E before and after applying color and location calibration.

Target model	Case of attack success	Basic attack			Color & 5 attack samples			Color & 10 attack samples		
		org	target	others	org	target	others	org	target	others
DNN_A	$Subj_3 \rightarrow Subj_4$	257	0	0	272	0	6	269	0	13
	$Subj_3 \rightarrow Subj_5$	266	0	0	–	–	–	277	0	8
	$Subj_4 \rightarrow Subj_6$	265	0	0	282	4	0	271	13	0
DNN_C	$Subj_0 \rightarrow Subj_1$	314	3	0	323	7	0	322	15	0

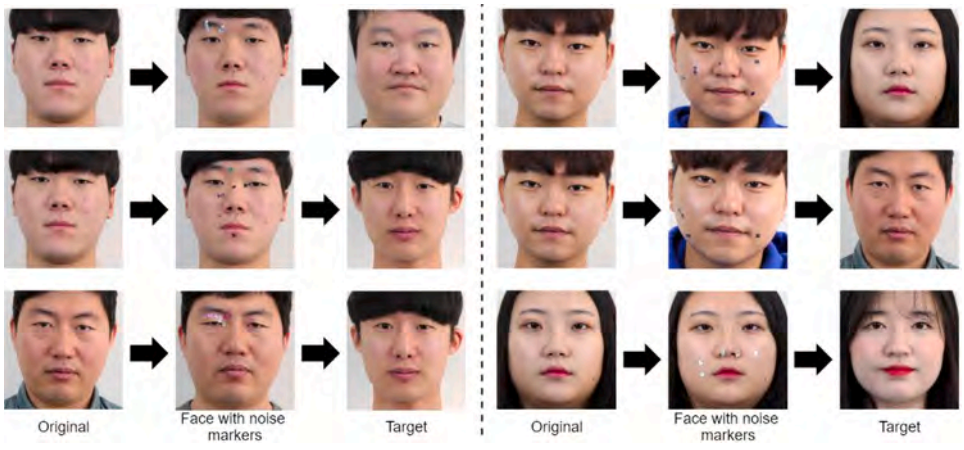

Fig. 7. Example of successful attacks by attaching noise markers on the face.

deceive FR models using face painting stickers that are more natural and have a wider modulation area than noise markers, and will try to attack commercial FR systems such as Face++ [43] and Amazon Rekognition [44].

(a) (b)

Fig. 8. Adversarial facial images generated by (a) a glasses attack and (b) our attack.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2020R1A2C1014813).

References

[1] Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process Mag 2012;82:82–97.

[2] Andor D, Alberti C, Weiss D, Severyn A, Presta A, Ganchev K, et al. Globally normalized transition-based neural networks. 2016, arXiv preprint arXiv:1603.06042.

[3] Krizhevsky A, Sustskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012, p. 1097–105.

[4] Taigman Y, Yang M, Ranzato M, Wolf L. Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 1701-08.

[5] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: The 26th British machine vision conference. 2015.

[6] Wang F, Cheng J, Liu W, Liu H. Additive margin softmax for face verification. IEEE Signal Process Lett 2018;25(7):926–30.

[7] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. In: International conference on learning representations. 2013.

[8] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: IEEE European symposium on security and privacy. 2016. p. 372-87.

[9] Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 2574-82.

[10] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: IEEE symposium on security and privacy. 2017. p. 39-57.

[11] Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 1625-34.

[12] Zhao Y, Zhu H, Liang R, Shen Q, Zhang S, Chen K. Seeing Seeing isn't believing: Towards more robust adversarial attack against real world object detectors. In: Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 2019. p. 1989-2004.

[13] Rozsa A, Günther M, Rudd EM, Boult TE. Facial attributes: Accuracy and adversarial robustness. Pattern Recognit Lett 2019;124:100–8.

[14] Dabouei A, Soleymani S, Dawson J, Nasrabadi N. Fast geometrically-perturbed adversarial Faces. In: IEEE winter conference on applications of computer vision. 2019. p. 1979-88.

[15] Zhou Z, Tang D, Wang X, Han W, Liu X, Zhang K. Invisible mask: Practical attacks on face recognition with infrared. 2018, arXiv preprint arXiv:1803.04683.

[16] Sharif M, Bhagavatula S, Bauer L, Reiter MK. 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. p. 1528-40.

[17] Zheng Y, Pal DK, Savvides M. Ring loss: Convex feature normalization for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 5089-97.

[18] Simonyan K, Zisserman A. Very deep convolution network for large-scale image recognition. In: International conference on learning representations. 2015.

[19] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 1-9.

[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770-8.

[21] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10, 000 classes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. p. 1891-8.

[22] Sun Y, Chen Y, Wang X, Tang X. Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. 2014, p. 1988–96.

[23] Sun Y, Wang X, Tang X. Deeply learned face representations are sparse, selective, and robust. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 2892-900.

[24] Sun Y, Liang D, Wang X, Tang X. Deepid3: Face recognition with very deep neural networks. 2015, arXiv preprint arXiv:1502.00873.

[25] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. p. 815-23.

[26] Wen Y, Zhang K, Li Z, Qiao Y. A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. 2016. p. 499-515.

[27] Liu W, Wen Y, Yu Z, Yang M. Large-margin softmax loss for convolutional neural networks. In: The 33rd international conference on machine learning. 2016.

[28] Liu W, Wen Y, Yu Z, Li M, Raj B, Song L. Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 212-20.

[29] Wang H, Wang Y, Zhou Z, Ji X, Li Z, Gong D, et al. Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 5265-74.

[30] Deng J, Guo J, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2019. p. 4690-99.

[31] Zhang X, Fang Z, Wen Y, Li Z, Qiao Y. Range loss for deep face recognition with long-tail. In: Proceedings of the international conference on computer vision. 2017. p. 5409-18.

[32] Wang F, Xiang X, Cheng J, Yuille AL. NormFace: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM international conference on Multimedia. 2017. p. 1041-9.

[33] Liu Y, Li H, Wang X. Rethinking feature discrimination and polymerization for large-scale recognition. 2017, arxiv preprint arXiv:1710.00870.

[34] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: International conference on learning representations. 2015.

[35] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. In: International conference on learning representations. 2017.

[36] Mirjalili V, Ross A. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In: IEEE international joint conference on biometrics. 2017. p. 564-73.

[37] Shen S, Furuta R, Yamasaki T, Aizawa K. Fooling neural networks in face attractiveness evaluation: Adversarial examples with high attractiveness score but lowsubjective score. In: 2017 IEEE third international conference on multimedia big data. 2017. p. 66-9.

[38] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Process Lett 2016;23(10):1499–503.

[39] He K, Zhang X, Ren S, Sun J. Identity mappings in deep residual networks. In: European conference on computer vision. 2016. p. 630-45.

[40] AI-Hub. K-Face. https://www.aihub.or.kr/aidata/73.

[41] Sandler M, Howard A, Menglong Z, Andrey Z, Liang-Chieh C. MobileNetV2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 4510-20.

[42] Szegedy C, Ioffe S, Vanhoucke V. Inception-v4, inception-resnet and the impact of residual connectionson learning. 2016, arXiv preprint arXiv:1602.07261.
[43] Megvii Inc. Face++. http://www.faceplusplus.com/.
[44] Amazon Web Services Inc. Amazon Rekognition. https://aws.amazon.com/rekognition/.

Gwonsang Ryu received the B.S. degree in applied mathematics from Kongju National University, South Korea, in 2016, and the M.S. degree in convergence science from Kongju National University, South Korea, in 2018. He is currently pursuing a Ph.D. degree in software convergence from Soongsil University, South Korea. His research interests include adversarial attacks, adversarial defense, user authentication, and anomaly detection.

Hosung Park received the B.S., M.S., and Ph.D degrees in computer engineering from Chungnam National University, Korea, in 2008, 2010, and 2014, respectively. He was a research associate at department of medical information, Kongju National University, Korea, from Aug. 2017 to Aug. 2020, and he is currently a professor with the department of cyber security and police, Busan University of Foreign Studies, South Korea. His research interests include information security and identity management.

Daeseon Choi received the B.S. degree in computer science from Dongguk University, South Korea, in 1995, the M.S. degree in computer science from the Pohang Institute of Science and Technology, South Korea, in 1997, and the Ph.D. degree in computer science from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2009. He was a professor with the department of medical information, Kongju National University, South Korea, from Sep. 2015 to Aug. 2020, and he is currently a professor with the department of software Soongsil University, South Korea. His research interests include identity management and information security.