IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 May 2023, accepted 21 May 2023, date of publication 24 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3279488

RESEARCH ARTICLE

Invisible Adversarial Attacks on Deep Learning-Based Face Recognition Models

CHIH-YANG LIN[1], (Senior Member, IEEE), FENG-JIE CHEN[2], HUI-FUANG NG[3], (Member, IEEE), AND WEI-YANG LIN[2,4], (Member, IEEE)

[1]Department of Mechanical Engineering, National Central University, Taoyuan 32001, Taiwan
[2]Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi 62102, Taiwan
[3]Department of Computer Science, University Tunku Abdul Rahman, Kampar 31900, Malaysia
[4]Advanced Institute of Manufacturing with High-Tech Innovations, National Chung Cheng University, Chiayi 62102, Taiwan

Corresponding authors: Hui-Fuang Ng (nghf@utar.edu.my) and Wei-Yang Lin (wylin@cs.ccu.edu.tw)

This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 111-2221-E-155-039-MY3 and Grant NSTC 110-2221-E-155-039-MY3.

ABSTRACT Deep learning technology has grown rapidly in recent years and achieved tremendous success in the field of computer vision. At present, many deep learning technologies have been applied in daily life, such as face recognition systems. However, as human life increasingly relies on deep neural networks, the potential harms of neural networks are being revealed, particularly in terms of deep neural network security. More and more studies have shown that existing deep learning-based face recognition models are vulnerable to attacks by adversarial samples, resulting in misjudgments that could have serious consequences. However, existing adversarial face images are rather easy to identify with the naked eye, so it is difficult for attackers to carry out attacks on face recognition systems in practice. This paper proposes a method for generating adversarial face images that are indistinguishable from the source images based on facial landmark detection and superpixel segmentation. First, the eyebrows, eyes, nose, and mouth regions are extracted from the face image using a facial landmark detection algorithm. Next, the superpixel segmentation algorithm is used to include the pixels neighboring the extracted facial landmarks with similar pixel values. Lastly, the segmented regions are used as masks to guide existing attack methods to insert adversarial noise within the masked areas. Experimental results show that our method can generate adversarial samples with high Structural Similarity Index Measure (SSIM) values at the cost of a small percentage of attack success rate. In addition, to simulate real-time physical attacks, printouts of the adversarial images generated by the proposed method are presented to the face recognition system via a camera and are still able to fool the face recognition model. Experimental results indicated that the proposed method can successfully perform adversarial attacks on face recognition systems in real-world scenarios.

INDEX TERMS Adversarial attack, deep learning, face recognition.

I. INTRODUCTION

In recent years, with the development of deep learning technology, more and more high-performance deep neural networks have been proposed by researchers. The design of network architecture has also become more complex and diverse. Face recognition is an important field in deep learning and computer vision. More and more products using face recognition models have emerged in our daily lives.

However, Szegedy et al. [15] first discovered that by adding perturbations over clean images, they were able to generate adversarial examples that could lead to model misjudgments. This is a serious flaw in the deep learning models. Since then, researchers have extensively explored adversarial attacks on deep learning models and have suggested an increasing number of attack strategies. For example, in 2019, the Real AI team at Tsinghua University in China used an

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar.

algorithm to generate glasses with special patterns. After actual testing, it was found that wearing such glasses could crack the face recognition systems of 19 Android phones on the market and could even complete the online authentication process for account registration on mobile banking APPs.

Nonetheless, existing adversarial sample generation techniques for face recognition models have limitations that restrict their practical applicability. It mainly depends on striking a balance between the attack success rate and the image quality of the adversarial samples. The attacks are typically carried out by adding unique patterns to the image of the face, which lowers the quality of the image. In this way, the attacks can be easily discovered through manual screening. Conversely, retaining the image's originality and quality will lead to a decrease in the attack success rate. As a result, many researchers are focusing on generating adversarial samples that are imperceptible to the naked eye and can successfully attack face recognition systems in a real-world setting.

Adding perturbation to the entire face image, or wearing specially designed adversarial glasses [14] or adversarial hats [9], will make the adversarial image look awkward and easy to detect. This study, on the other hand, proposed introducing adversarial perturbations only in regions with complex textures, such as facial features in the face image. There are two benefits to doing this. First, hiding adversarial noises in highly textured regions will make them difficult to discover visually. Second, modifying facial features is the most effective way to affect the performance of the face recognition models. Our method first uses facial landmark detection and a superpixel segmentation algorithm to generate a mask on the facial feature regions, and then applies the existing gradient-based attack method to inject perturbations into the masked areas. Experiments show that the proposed method is able to generate adversarial samples that are more imperceptible visually while maintaining attack efficiency on face recognition models.

The main contributions of this paper are as follows:

1) This study is the first to combine facial landmark detection with superpixel segmentation to effectively generate a mask that covers the facial feature regions in the face image.
2) Masked facial feature regions are used as attack points, which greatly improve the structural similarity between an adversarial sample and its source image at the cost of a small percentage of attack success.
3) Our method can be used together with various attack methods. Compared with the original attack method, our method is more visually concealed and thus can be applied in real-world scenarios.

II. RELATED WORK

A. ADVERSARIAL ATTACKS

In recent years, many adversarial attack algorithms have been proposed and have demonstrated effective attacks on deep learning models. Khamaiseh et al. [8] categorize existing attack algorithms into white-box attacks and black-box attacks. The former needs to know important information about the target model (weight parameters, training data, etc.), while the latter does not require such information. Although black-box attacks do not require information about the model, their success rate is generally much lower than that of white-box attacks. In this paper, three classical white-box attack methods are used as a baseline for experiments, which are described below.

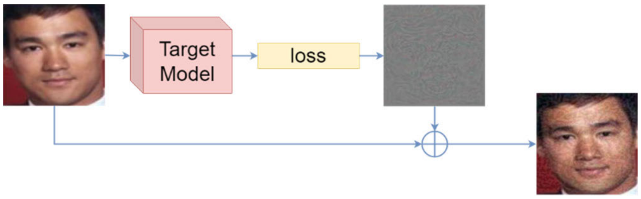

FIGURE 1. Flow diagram of FGSM [5].

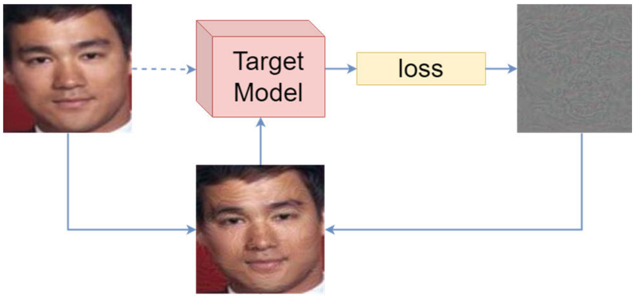

FIGURE 2. Flow diagram of BIM [10].

Fast Gradient Sign Method [5], or FGSM for short, is the first proposed adversarial attack method by Goodfellow et al. [5]. FGSM exploits the gradients of a neural network to build an adversarial image. As shown in Figure 1, given an input image, FGSM computes the gradients of the loss with respect to the image. It then uses the sign of the gradients as perturbations to create a new image (the adversarial image) that maximizes the loss. The Fast Gradient Sign Method can be expressed by the following equation:

$$x' = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y)) \tag{1}$$

where x is the original input image, y is the ground truth label, and θ represents model parameters. $J(\theta, x, y)$denotes the lost function and $\nabla_x J(\theta, x, y)$ is the gradient of the lost function. ϵ is a small value to multiply the signed gradients to ensure that the perturbations are small, and x' is the output adversarial image. FGSM is a typical one-step attack method. This method is simple and fast, but the success rate is not high; it achieved only a 63% to 69% success rate on ImageNet.

The Basic Iterative Method (BIM) [10] is an improved method based on FGSM. As shown in Figure 2, BIM incorporates iteration steps into the FGSM process. As a result, BIM can be seen as an iterative version of FGSM, also known

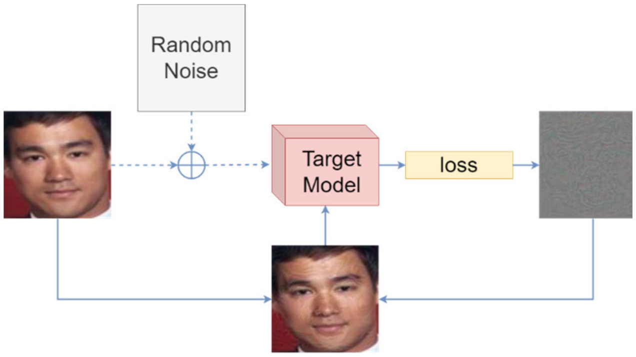

FIGURE 3. Flow diagram of PGD [11].

as I-FGSM. BIM is expressed as:

$$X_0^{adv} = x \tag{2}$$

$$X_{N+1}^{adv} = Clip_{x,\epsilon} X_N^{adv} + \alpha \cdot sign(\nabla_x J(\theta, X_N^{adv}, y))\} \tag{3}$$

where X_0^{adv} is the initial adversarial sample, defaulting to the input image x. X_N^{adv} represents the adversarial sample at iteration N. α is the multiplier for controlling the perturbation magnitude. X_{N+1}^{adv} is the current adversarial sample obtained by adding updated perturbations to previous sample X_N^{adv} and clipping at the x$\pm\epsilon$ range. Compared with FGSM, BIM generally has a higher attack success rate, and the generated adversarial samples look smoother, as seen in Figure 2.

Projected Gradient Descent (PGD) [11] is another improved method based on FGSM. PGD is also referred to as K-FGSM, where K is the number of iterations. Both PGD and BIM are iterative versions of FGSM. As shown in Figure 3, the main difference between the two is that PGD uses more iterations and adds random disturbances to the input during initialization. The formulation of PGD is shown below:

$$x^0 = x + noise \tag{4}$$

$$x^{t+1} = \prod\nolimits_{x \pm S} (x^t + \alpha \cdot sign(\nabla_x J(\theta, x^t, y))) \tag{5}$$

x^t represents the generated adversarial image at iteration t. In each iteration, the new adversarial image is obtained by adding updated perturbations to the image from previous iteration and clipping at the $x \pm s$ range.

B. ADVERSARIAL ATTACKS ON FACE RECOGNITION

With the increase in attack algorithms, researchers began to design new attack methods on face recognition models to evaluate their robustness. Since face recognition models are widely used in practice, it is common for researchers to develop attack methods that can be applied in real-world scenarios. The following introduces several well-known adversarial attacks on face recognition.

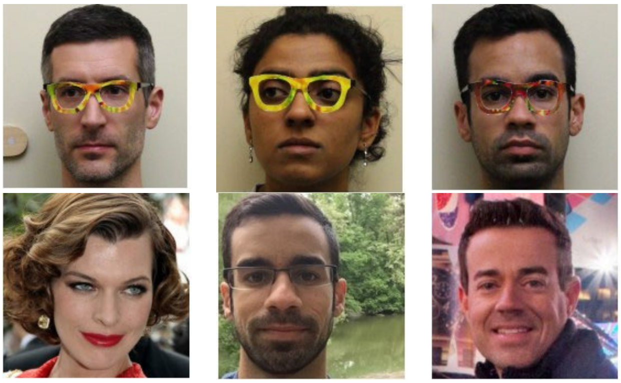

FIGURE 4. Successful cases of the Adv-Glasses attack [14]. The top row are the original face images plus Adv-Glasses and the bottom row are the respective wrongly recognized targets.

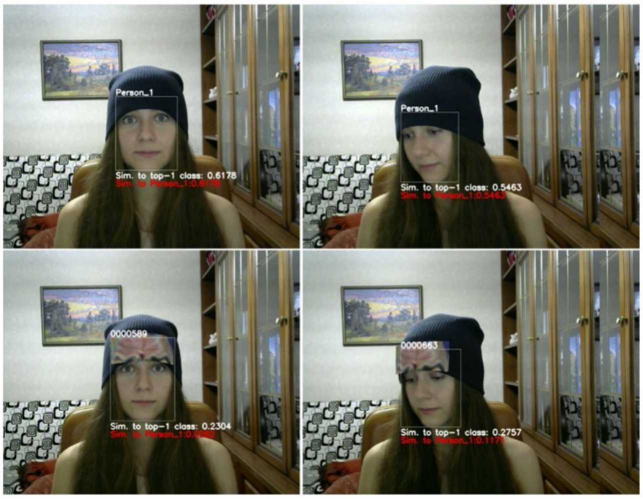

FIGURE 5. Successful cases of the Adv-Hat attack [9].

1) ADVERSARIAL GLASSES

Sharif et al. [14] proposed a method to generate a perturbed glasses-style patch, named Adv-Glasses, and then add this patch to the face image to perform the attack on face recognition systems. Considering real-world application scenarios, the authors impose several conditions to make the disturbance look smoother. Figure 4 shows a few success cases of the Adv-Classes attack. According to the authors, this method achieved a 100% attack success rate under the white-box, untargeted attacks in the digital world and can also achieve a certain success rate under targeted attacks. However, the dataset used in the paper contains only 10 people, so the difficulty of the attack is relatively low. The authors also mentioned that the success rate of the attack will gradually decrease as the number of people increases.

2) ADVERSARIAL HAT

Using a similar principle as the Adv-Glasses method, Komkov et al. [9] designed the adversarial hat (Adv-Hat) attack method targeting the Arcface-based Face ID system. In this method, a large disturbance patch is created and placed on the forehead as a hat. As shown in Figure 5, with the patch, the similarity score for the adversarial sample dropped from

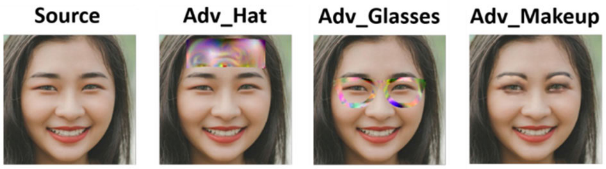

FIGURE 6. Comparison of adversarial samples generated by different attack methods [18].

about 59% (left column, frontal view) to 43% (right column, side view) compared to the original image, causing wrong classification by the face recognition model. The authors also reported that the attack can still be successful after printing the adversarial samples on paper and capturing the images through the camera, which indicates that the Adv-Hat method can be applied in real-world applications. However, the drawback of this method is that the generated patch is too abrupt to the naked eye, and the success rate of the attack will decrease when the view angle is shifted.

3) ADVERSARIAL MAKEUP

Yin et al. [18] proposed the adversarial makeup (Adv-Makeup) method, which synthesizes perturbations as realistic eye shadow over the orbital region of the source face image. As shown in Figure 6, the images generated by Adv-Makeup are more natural than those generated by other methods. However, there are limitations to implementing such methods in the real-world scenario. For example, pixel intensity values might change when outputting adversarial samples to the physical world, which will affect the integrity of the perturbation and, in turn, will lead to a decrease in the attack success rate. Another issue is that although Adv-Makeup generated images are more realistic compared to previous methods, they can still be easily detected by the naked eye. Therefore, these methods are currently mainly applied in the digital world.

C. 3D ADVERSARIAL ATTACKS

Researchers have recently begun to realize that 3D adversarial attacks are more reminiscent of actual attack scenarios. An innovative and effective method for 3D adversarial attacks that is applicable to a variety of 3D representations was introduced by Zhang et al. [20]. Given the growing usage of 3D models in a wide range of applications and the potential security threats posed by 3D adversarial attacks, this research is crucial and timely. Chen et al. [1] proposed a dual-stream architecture for identifying presentation attacks in 3D mask faces. It consists of two streams: a color stream and a depth stream. The color stream processes RGB images captured from a standard camera, while the depth stream processes depth maps generated from a structured-light 3D scanner. The two streams are combined using a fusion layer to produce a final classification result.

III. METHODS

The proposed attack method consists of two parts: the Mask Generation Block and the Adversarial Attack Block, as shown in Figure 7. The mask generation block uses a variety of algorithms to segment regions with complex texture features in a face image. The segmented regions contain facial landmarks that are important for the face recognition model to perform recognition, and therefore they are also optimal places for hiding adversarial attack noise in the face image. The segmented regions form a mask to guide the adversarial attack block to insert adversarial noise into the masked areas using existing attack methods. Extra loss functions are incorporated into the framework to make the generated adversarial samples smoother to ensure that the adversarial samples can successfully attack the target face recognition model in both digital and physical scenarios, and that the hidden adversarial noise is unnoticeable to the naked eye. This section describes each component of the proposed framework in detail.

A. MASK GENERATION BLOCK

Zagoruyko and Komodakis [19] study where the face recognition model is focusing on in an image by summing the attention maps output by each of the convolutional layers in the model. As shown in Figure 8, it can be seen clearly from the mid-level and high-level attention maps that the model is mostly focusing on the facial landmark regions. Therefore, adding perturbations to the facial landmark regions will be an effective way to influence the face recognition model to produce erroneous results.

In the proposed mask generation block, facial landmark detection and superpixel segmentation are used to segment out facial landmark regions in the input face image. The extracted regions will be used to form a foreground mask for the adversarial attack block to insert adversarial noise into the masked areas.

1) FACIAL LANDMARK DETECTION

In previous face recognition models, facial landmark detection was mainly used to find the position of the face in the image to assist the face recognition model in performing a recognition task. Kazemi et al. [7] proposed the well-known Ensemble of Regression Tress algorithm to locate 68 landmarks in a face image, including the eyebrows, eyes, nose, mouth, and jawline of the face, as shown in Figure 9. Inspired by this, Kazemi's method is first used to locate the facial landmarks in the face image. The detected landmark points are then linked together to generate an initial foreground mask, as shown in Figure 10.

2) SUPERPIXEL SEGMENTATION

The initial foreground mask generated by the facial landmark detection algorithm contains only the primal shapes of the facial features. It misses the fine details surrounding the facial landmark regions, which are important for generating stealthy adversarial samples. To resolve this issue, the superpixel segmentation algorithm is applied to incorporate the edge pixels surrounding the facial landmark regions. Superpixel segmentation groups neighboring pixels into clusters with

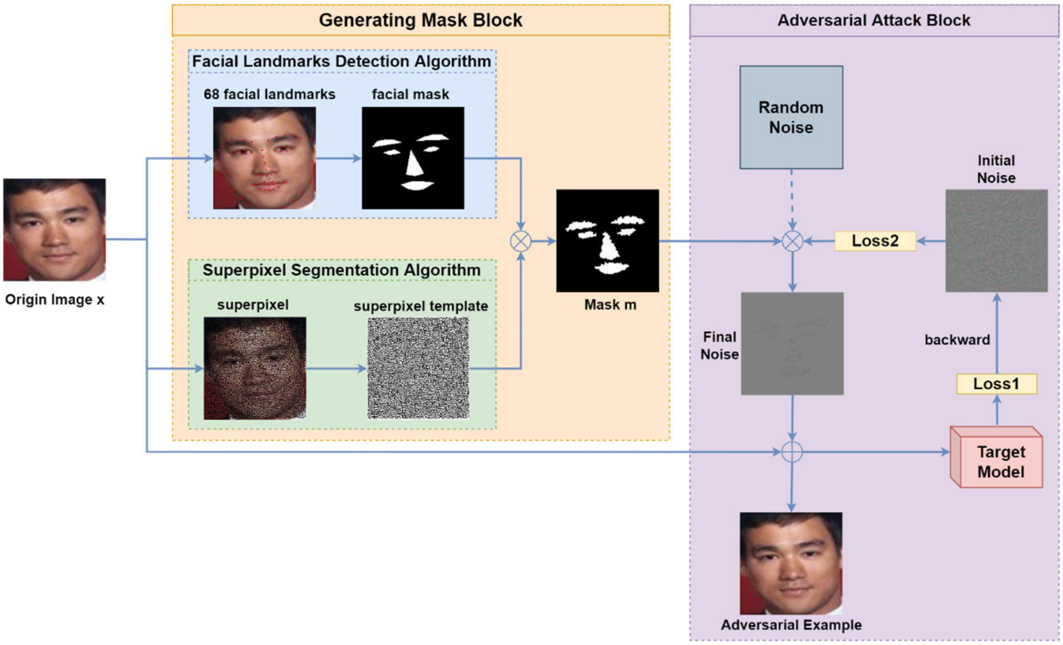

FIGURE 7. Overall framework of the proposed method.

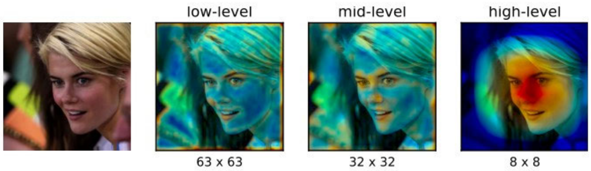

FIGURE 8. The focus areas of the face recognition model at different convolution layers. Reddish colors indicate higher levels of attention [19].

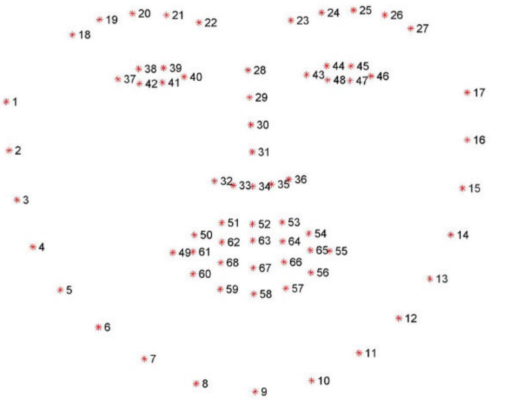

FIGURE 9. Positions of 68 facial landmarks [7].

similar properties (color, texture, etc.). The superpixel segmentation algorithm divides the image into areas with similar pixel values to form the superpixel template, as shown in Figure 11. To generate the final mask, the superpixel template and the initial facial landmark mask are combined.

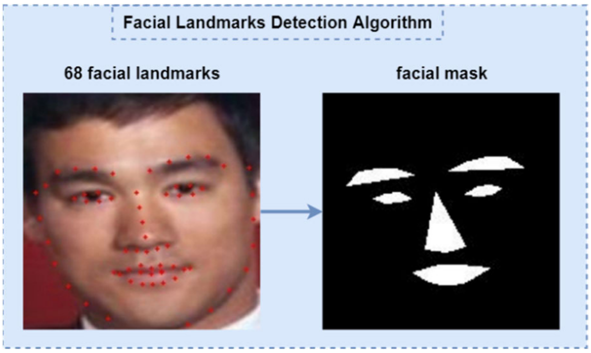

FIGURE 10. Initial mask generated using the facial landmark detection algorithm.

A superpixel is defined as foreground if it intersects with any foreground regions in the initial mask. As shown in Figure 12, the superpixel template recovered the missing edge pixels in the original mask, making the final mask look more natural.

B. ADVERSARIAL ATTACK BLOCK

The mask generated from the mask generation block will be used to guide the adversarial attack block to insert adversarial noise in the foreground regions covering the facial landmarks. The flow diagram in Figure 13 shows the steps involved in the adversarial attack block based on the PGD … [11] attack algorithm. In the initial iteration, random noises are added to the masked foreground regions in the original face image, and the result image is input to the face recognition model to get the gradients of the Cross Entropy Loss (Loss1) through backpropagation. The gradients are then passed through another

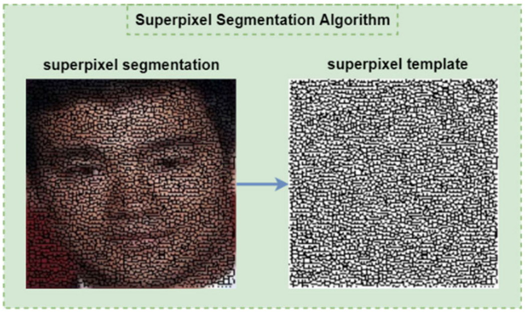

FIGURE 11. Superpixel template generated using a superpixel segmentation algorithm.

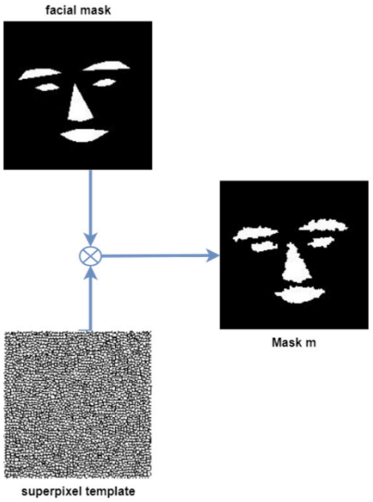

FIGURE 12. Final mask after combining the initial mask and the superpixel template.

loss function, the Total Variation Loss (Loss2), to smooth the perturbations. More details of the lost functions are given in the following section. Finally, the smoothed signed gradients are cropped at the masked regions and inserted into the input image to generate the adversarial image. This process is then repeated, and the final adversarial sample is obtained after completing the specified number of iterations. The formulation of the adversarial attack block is expressed as follows:

$$x_0^{adv} = x + Crop_m(r) \tag{6}$$

$$x_{i+1}^{adv} = Clip_{x,\epsilon} x_i^{adv} + Crop_m(\alpha \cdot sign(\nabla_x L(x_i^{adv}, y_{true})))\} \tag{7}$$

In Eq. 6, r represents the initial random noise and $Crop_m$ function indicates that the noises or perturbations are constrained within the foreground regions obtained from the mask generation block. x_0^{adv} is the initial adversarial sample at iteration 0, defaulting to the initial face image x plus the cropped random noise. In Eq. 7, $L(x_i^{adv}, y_{true})$ denotes the lost function where x_i^{adv} represents the adversarial sample at iteration i and y_{true} is the ground truth label. $\nabla_x L(x_i^{adv}, y_{true})$ is the gradient of the lost function and α is a multiplier to

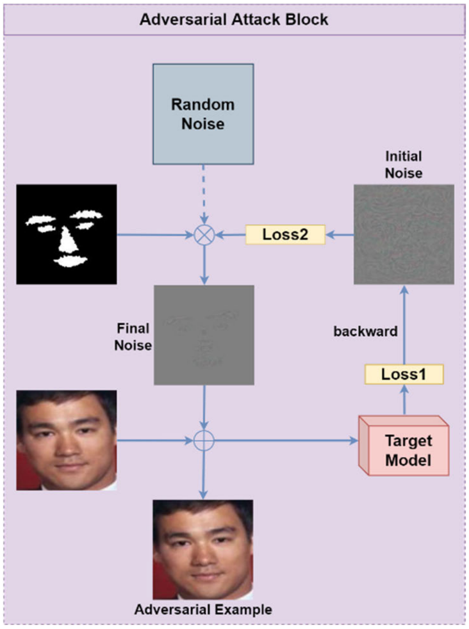

FIGURE 13. Adversarial attack block flow diagram.

ensure that the perturbation values are small. In the similar manner, the updated perturbations are cropped within the masked regions. The new adversarial sample x_{i+1}^{adv} is obtained by adding the cropped perturbation to the current sample x_i^{adv} and clipping at the x±ϵ range.

C. LOST FUNCTIONS

There are two lost functions used in adversarial attack generation: the cross-entropy loss (L_{CE}) and the total variation loss (L_{TV}). Cross-entropy loss is mainly used to find the gradients required for generating perturbations. The purpose of the total variation loss is to make the generated adversarial sample smoother, which allows it to better retain the added perturbations when outputting to the real world.

1) CROSS-ENTROPY LOSS

Cross-entropy loss [2] is one of the loss functions used in training multi-class classification models. In classification model training, cross-entropy loss is computed using the prediction score obtained after inputting the image into the model. The smaller the loss value, the higher the probability that the input image belongs to the correct category. In contrast, for an adversarial attack, the extra perturbations should make the model output a larger loss value, indicating that the model is more likely to make an incorrect prediction and thus the attack is successful. The cross-entropy loss (L_{CE}) formula is expressed as:

$$L_{CE}(x, y) = -\log\left(\frac{\exp(x[y])}{\sum_j \exp(x[j])}\right) = -x[y] + \log(\sum_j \exp(x[j])) \tag{8}$$

where $x[j]$ is the prediction score corresponding to each category j after the image is input into the model and y is the ground truth label.

2) TOTAL VARIATION LOSS

During a physical adversarial attack, such as photocopying an adversarial face sample and retaking the face image via a camera, the noise suppression function of the camera will weaken the adversarial perturbations originally implanted in the image and cause the attack to fail. To mitigate the problem, a Total Variation Loss [12] function is incorporated in the process of generating adversarial samples to make the adversarial samples smoother so that they are less susceptible to noise suppression. Total variation loss enhances the smoothness of an image by reducing the differences between adjacent pixel values. The total variation loss (L_{TV}) is expressed as:

$$L_{TV}(r) = \sum\nolimits_{i,j} \left(\left(r_{i,j} - r_{i+1,j}\right)^2 + \left(r_{i,j} - r_{i,j+1}\right)^2 \right)^{\frac{1}{2}} \tag{9}$$

where r is the input image, $r_{i,j}$ is the pixel value at coordinate (i, j) in r. The smaller the L_{TV} value, the closer the adjacent pixel values are, and therefore, the smoother the image.

IV. EXPERIMENTS

In this section, the performance of the proposed adversarial attack method on face recognition models is evaluated. All experiments utilize the same testing environment and hyperparameter settings.

A. EXPERIMENTAL SETUP

The public CASIA-WebFace [17] face dataset is used in the experiments to evaluate the proposed adversarial attack method on various face recognition models. The dataset contains a total of 494,414 images of 10,575 celebrities, the largest number of face categories currently available. After excluding low-quality images, a total of 296,624 images were used in the experiments.

The FaceNet [13], IR152, IRSE50, and ArcFace [4] face recognition models are chosen for the experiments. Among them, the FaceNet model pre-trained on CASIA-WebFace is used to generate the adversarial samples and to conduct white-box attack experiments on FaceNet. The same adversarial samples are used to perform black-box attack experiments on the other three models.

Three attack methods, FGSM [5], BIM [10], and PGD [11], are used as benchmarks to assess the attack performance enhancements of the proposed mask generation block and adversarial attack block.

B. EVALUATION METRICS

The effectiveness of the proposed method is evaluated using two performance metrics: Attack Success Rate (ASR) [3] and Structural Similarity Index Measure (SSIM) [16].

1) ATTACK SUCCESS RATE (ASR)

ASR is used to evaluate the overall attack success rate of the generated adversarial samples on face recognition models. The value is between 0 and 100%. The closer to 100%, the higher the attack success rate of the adversarial samples on the target model, and vice versa. The equation of ASR is as follows:

$$ASR = \frac{\sum_i^N 1_\tau (\cos\left[F\left(I_s^i\right), F\left(\hat{I}_s^i\right)\right] < \tau)}{N} \times 100\% \tag{10}$$

In the above equation, I_s^i represents the original face image and $\hat{I}_s^i$ is the corresponding adversarial image. F represents the feature vector produced by the face recognition model from the input face image. The *cos* function computes the cosine similarity between the two face feature vectors. When the cosine similarity value approaches 1, it indicates that the two images are more similar, and vice versa. τ is a predefined threshold value. Different face recognition models require different τ values. For example, the τ values used for each of the models are 0.167 (IR152), 0.241 (IRSE50), 0.302 (ArcFace), and 0.409 (FaceNet). Next, 1_τ is an indicator function which returns 1 when the condition is *true* and returns 0 otherwise. Lastly, N denotes the total number of adversarial examples. In short, ASR computes the percentage of adversarial images that are different from their original face images from the perspective of a face recognition model.

2) STRUCTURAL SIMILARITY INDEX MEASURE (SSIM)

Our method focuses on generating adversarial face images that are indistinguishable from the source images. The SSIM index is adopted to evaluate the visual similarity between an adversarial sample and its original image. SSIM takes into account the brightness, contrast, and structure of an image during comparison. That is, the SSIM index is closer to human vision when comparing images. The values of SSIM range between 0 and 1, with 1 indicating the two images are identical. SSIM is expressed as:

$$SSIM\left(x, x'\right) = l(x, x') \times c(x, x') \times s(x, x') \tag{11}$$

where

$$l\left(x, x'\right) = \frac{2u_x u_{x'} + C_1}{u_x^2 + u_{x'}^2 + C_1}$$
$$c\left(x, x'\right) = \frac{2\sigma_x \sigma_{x'} + C_2}{\sigma_x^2 + \sigma_{x'}^2 + C_2}$$
$$s\left(x, x'\right) = \frac{\sigma_{xx'} + C_3}{\sigma_x \sigma_{x'} + C_3}$$

From the equation, $l\left(x, x'\right)$, $c\left(x, x'\right)$, and $s\left(x, x'\right)$ compare the brightness, contrast, and structure between x and x', respectively. u_x and σ_x are the average and standard deviation of x, and $\sigma_{xx'}$ is the covariance of x and x'. C_1, C_2, and C_3 are predefined constants.

TABLE 1. Attack success rate (%) results of various attack methods on the CASIA-WebFace dataset.

Attack Methods	FaceNet (White-box)	IR152 (Black-box)	IRSE50 (Black-box)	ArcFace (Black-box)
FGSM	90.72%	1.40%	16.62%	25.30%
FGSM+mask(ours)	61.44%	0.38%	14.86%	24.64%
BIM	93.11%	4.66%	22.56%	34.54%
BIM+mask(ours)	90.72%	3.20%	20.78%	32.30%
PGD	100.0%	5.92%	25.62%	38.62%
PGD+mask(ours)	90.56%	4.38%	23.94%	36.70%

TABLE 2. SSIM between original images and adversarial images generated by various attack methods on the CASIA-WebFace dataset.

Attack Methods	SSIM
FGSM	0.5564
FGSM+mask(ours)	**0.9714**
BIM	0.8523
BIM+mask(ours)	**0.983**
PGD	0.835
PGD+mask(ours)	**0.979**

C. RESULTS

This section presents the experimental results of applying the proposed method to both the digital-world attacks and the physical-world attacks.

1) DIGITAL-WORLD ATTACKS

As mentioned in IV.A, the FaceNet model pre-trained on CASIA-WebFace is used as the baseline model to generate the adversarial samples, and the samples are used to simulate the white-box attack on the FaceNet itself. The same adversarial samples are then used to simulate black-box attacks on IR152, IRSE50, and ArcFace, respectively. In the experiments, the perturbation limit parameter ϵ is set at 20/255 since this value produced the best overall results and was also shown to produce successful attacks in the real world.

Table 1 shows the attack success rates of different attack methods on the CASIA-WebFace dataset, including the results of white-box attacks and black-box attacks. It can be seen that the attack success rates of the adversarial samples generated after applying our mask decreased slightly compared to the results obtained without the mask. However, the range of decrease is minor, indicating that the proposed mask is effective in most images. The outcome is somewhat predictable because our method shifted the perturbations that were previously distributed across the entire face image onto the facial landmark regions, limiting the total number of perturbations that can be implanted in the face image and thus weakening the attack strength. However, our method improved the smoothness of the adversarial images, making them more natural-looking to the naked eye. Table 2 shows the SSIM values between the original face images and the respective adversarial images generated from different attack methods. It can be seen that the SSIM values for the three attack methods have increased significantly after applying the proposed mask. The visualization of the adversarial faces generated by the original attack methods and the proposed method in Figure 14 clearly shows that the image quality has been significantly improved after applying the mask, and the perturbation noise implanted in the image is almost imperceptible.

2) PHYSICAL-WORLD ATTACK

Compared with digital-world attacks, physical-world attacks are more in line with real-world applications. To simulate the physical-world attack, we output the digital adversarial images by printing them on paper and then capture the photos through a scanner. A physical-world attack example is shown in Figure 15. The image on the left is the original photo of Bruce Lee, and the image on the right is the printed version of the adversarial face. Visually, there is almost no difference between the two images, but the FaceNet model still misclassified the adversarial face, indicating that our attack method can be used in a real-world attack scenario.

The reason that the proposed method is able to maintain image quality while at the same time performing a successful attack is that the adversarial perturbations are strategically implanted in the areas that are most critical to the face recognition models, the facial landmark regions. A small amount of perturbation in these areas is enough to cause the face recognition model to make mistakes. Another reason is that hiding adversarial noises in highly textured facial landmark regions makes them difficult to discover visually, thus maintaining the image quality. In addition, since total variation loss is incorporated in generating the adversarial samples, the generated adversarial samples are smoother so that most of the added perturbations can be retained when exporting them to the real world.

For simulating real-time physical attacks, a camera is set up to capture and perform real-time face recognition from printouts of the original face and the adversarial face image. Figure 16 shows the real-time test results. The adversarial sample printout generated by the proposed method can deceive the FaceNet model into making the incorrect prediction. The results again show that our method is effective in real-world applications.

Additional experiments were carried out to examine the effects of external factors such as lighting conditions, camera viewing angles, and distance from the camera on the real-time physical attack performance. Three lighting conditions: low lighting, normal lighting, and strong lighting, were selected for the lighting test. Figure 17 shows the adversarial images captured by a camera under the three lighting conditions and their respective classification results by FaceNet. It can be seen that adversarial images captured under low lighting condition (left image in Figure 17) failed to deceive the FaceNet model. One possible explanation is that when the lighting is too low, the camera is not able to fully acquire the perturbations embedded in the adversarial photo, causing the attack to fail. However, under normal and strong lighting conditions

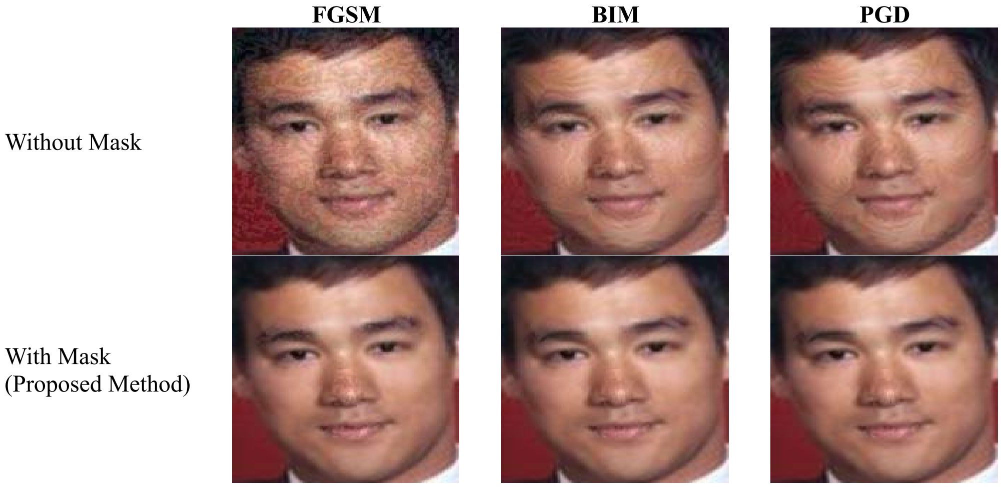

FIGURE 14. Visualization of adversarial faces generated by the original attack methods and the proposed method.

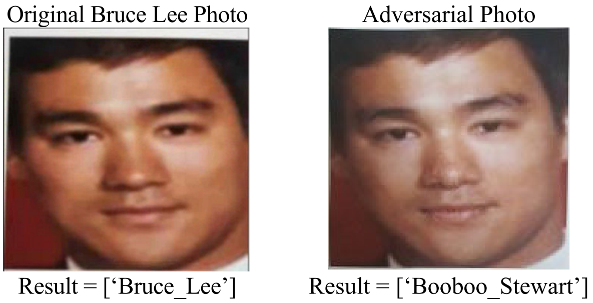

FIGURE 15. Physical-world attack examples. The image on the left is the printout of the original Bruce Lee photo, and FaceNet classified it correctly. The image on the right is the printout of the adversarial face generated by our method, and FaceNet misclassified the face as Booboo Stewart.

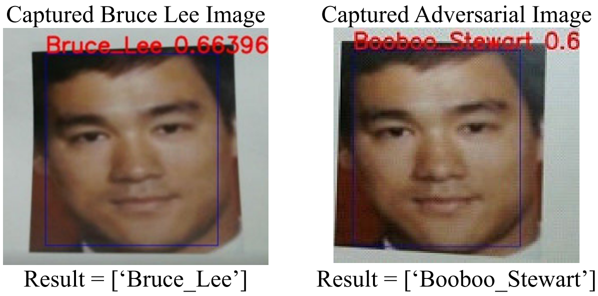

FIGURE 16. Real-time physical attack examples. On the left is a real-time image captured from a printout of Bruce Lee's photograph. The image on the right is captured from the printout of the adversarial face generated by our method, and FaceNet misclassified the face.

(middle and right images), the camera is able to pick up enough perturbation signals in the adversarial samples to fool the FaceNet model.

To simulate the effect of camera viewing angles on physical adversarial attacks, the camera was tilted 30 degrees in the upward, downward, left, and right directions while taking the image from the printout containing the adversarial sample.

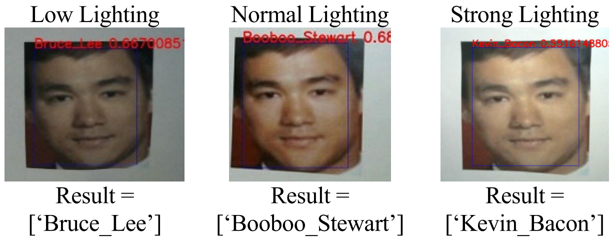

FIGURE 17. Real-time physical attack examples under three different lighting conditions. FaceNet misclassified the captured adversarial images under normal and strong lightings, but classified correctly for image captured under low lighting.

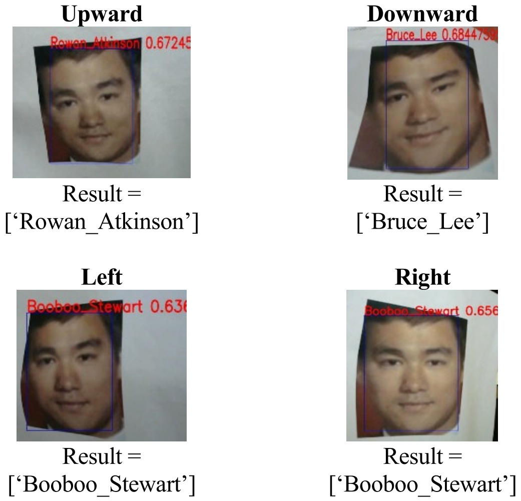

FIGURE 18. Real-time physical attack examples under different camera viewing angles. FaceNet misclassified the captured adversarial images in most cases except for the downward angle.

As shown in Figure 18, in most cases, the captured adversarial images are able to deceive the FaceNet model, except for the image captured at a downward angle. The results indicate that the proposed adversarial attack is generally robust against

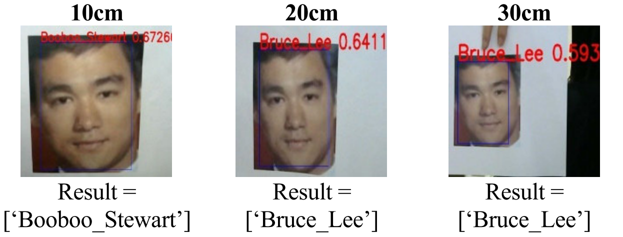

FIGURE 19. Real-time physical attack examples under different distances from the camera. The adversarial images failed to deceive FaceNet when they are captured at a distance far away from the camera.

TABLE 3. Attack success rate (%) of the proposed adversarial attack method against the RandomNoise_inpaint [21] and ComDefend [6] defense mechanisms.

Attack Method	No defense	RandomNoise -Inpaint [21]	ComDefend [6]
FGSM	90.72%	95.56 %	79.68 %
FGSM+mask(ours)	61.44%	95.91 %	59.20 %
BIM	93.11%	98.98 %	88.92 %
BIM+mask(ours)	90.72%	98.47 %	84.75 %
PGD	100.00%	99.91 %	100.00 %
PGD+mask(ours)	90.56%	98.55 %	85.26 %

changes in view angle. However, if the camera's view angle causes an excessive amount of distortion in the acquired image, the method may not work.

Lastly, to test the camera distance effect, the printout is placed at 10cm (normal distance), 20cm, and 30cm away from the camera. The images taken at the three camera distances and their classification results by FaceNet are shown in Figure 19. The results show that the adversarial sample failed to fool the FaceNet model when the distance of the sample to the camera was more than 20cm. As the distance to the camera increases, the captured image gets smaller, and thus fewer perturbations will be acquired. As a result, the chance of a successful attack will be lower.

In summary, the proposed method is generally robust against external factors such as lighting conditions, camera viewing angles, and distance from the camera. However, its physical attack performance might degrade if there are excessive amounts of distortion in the acquired images.

3) AGAINST ADVERSARIAL DEFENSES

To evaluate the robustness of the proposed attack method against different adversarial defense mechanisms, two adversarial defense methods, RandomNoise-Inpaint [21] and ComDefend [6], were tested. Table 3 shows the attack success rate of the proposed adversarial attack method against the two defense mechanisms. Our attack method is able to maintain high attack success rates, in some cases, even achieve a higher success rate against the RandomNoise-Inpaint defense mechanism. The ComDefend mechanism is more effective in defending the proposed attack, but its efficacy is limited. The attack success rates of the proposed method drop slightly from -2.24% (FGSM+mask) to -5.97% (BIM+mask). Therefore, the results indicate that the proposed adversarial attack is robust against different adversarial defense mechanisms.

V. CONCLUSION

This paper proposes a mask generation method based on facial landmark detection and superpixel segmentation for improving existing adversarial attack methods on face recognition models, both digital and physical. The proposed mask guides existing attack methods to insert adversarial noise at the facial landmark regions within the face image, making the attack effective and invisible to human vision. A smoothing loss is used in the generating process to make the adversarial samples suitable for real-world applications. Experimental results indicate that the proposed method is effective against both digital-world and physical-world attacks on face recognition systems.

REFERENCES

[1] S. Chen, T. Yao, K. Zhang, Y. Chen, K. Sun, S. Ding, J. Li, F. Huang, and R. Ji, "A dual-stream framework for 3D mask face presentation attack detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 834–841.

[2] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[3] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces: Adversarial face synthesis," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.

[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[6] X. Jia, X. Wei, X. Cao, and H. Foroosh, "ComDefend: An efficient image compression model to defend adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6077–6085.

[7] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[8] S. Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino, and H. W. Alomari, "Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification," *IEEE Access*, vol. 10, pp. 102266–102291, 2022.

[9] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace face ID system," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 819–826.

[10] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*. Boca Raton, FL, USA: CRC Press, 2018, pp. 99–112.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[12] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[13] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[14] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2016, pp. 1528–1540.

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[17] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[18] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," 2021, *arXiv:2105.03162*.

[19] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*.

[20] J. Zhang, L. Chen, B. Liu, B. Ouyang, Q. Xie, J. Zhu, W. Li, and Y. Meng, "3D adversarial attacks beyond point cloud," *Inf. Sci.*, vol. 633, pp. 491–503, Jul. 2023.

[21] F. Zuo and Q. Zeng, "Exploiting the sensitivity of l2 adversarial examples to erase-and-restore," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, May 2021, pp. 40–51.

CHIH-YANG LIN (Senior Member, IEEE) is currently with the Department of Mechanical Engineering, National Central University, Taoyuan, Taiwan. Previously, he was the Dean of the International Academy, the Chief of Global Affairs Office, and a member of the Department of Electrical Engineering, Yuan-Ze University, Taoyuan. He has been recognized as an IET Fellow and has contributed to over 200 papers that have been featured in a wide range of international conferences and journals. His research interests include computer vision, machine learning, deep learning, image processing, big data analysis, and the design of surveillance systems. He received the Best Paper Awards from the Pacific-Rim Conference on Multimedia (PCM), in 2008, the Best Paper Awards from the IPPR CVGIP Conference, in 2009, 2013, and 2019, the Best Paper Award from IVIC'19, and the Best Paper Award from BCWSP 2020. He also has served in several leadership positions for various international conferences, taking on responsibilities, such as the Program Chair, the Session Chair, the Publication Chair, the Publicity Chair, and a Workshop Organizer for events, such as AHFE, ICCE, ACCV, IEEE Multimedia Big Data, ACM IH&MMSec, APSIPA, and CVGIP.

FENG-JIE CHEN received the B.S. and M.S. degrees from National Chung Cheng University (CCU), Chiayi, Taiwan. His research interests include computer vision, image processing, deep learning, and adversarial attacks.

HUI-FUANG NG (Member, IEEE) received the Ph.D. degree in biosystems and agricultural engineering from the University of Minnesota, USA, in 1996. He was a Software Engineer with PPT Vision Inc., Minnesota, USA, from 1996 to 2003. He joined the Department of Computer and Information Science, Asia University, Taiwan, from 2003 to 2013. He was also the Chairperson of the Centre for IoT and Big Data, University Tunku Abdul Rahman (UTAR), Malaysia, from 2019 to 2021, where he is currently an Associate Professor with the Department of Computer Science. His research interests include image processing, computer vision, and machine learning.

WEI-YANG LIN (Member, IEEE) received the B.S.E.E. degree from National Sun Yat-sen University, Taiwan, in 1994, and the M.S.E.E. and Ph.D. degrees from the University of Wisconsin–Madison, in 2004 and 2006, respectively. Since 2006, he has been with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, where he is currently a Professor. His research interests include computer vision, biometric authentication, and multimedia signal processing.