DOI: 10.19884/j.1672-5220.202407001

Incomplete Physical Adversarial Attack on Face Recognition

HU Weitao[1], XU Wujun[1, 2*]
1. College of Information Science and Technology, Donghua University, Shanghai 201620, China
2. Engineering Research Center of Digitalized Textile and Fashion Technology, Donghua University, Shanghai 201620, China

Abstract: In recent work, adversarial stickers are widely used to attack face recognition (FR) systems in the physical world. However, it is difficult to evaluate the performance of physical attacks because of the lack of volunteers in the experiment. In this paper, a simple attack method called incomplete physical adversarial attack (IPAA) is proposed to simulate physical attacks. Different from the process of physical attacks, when an IPAA is conducted, a photo of the adversarial sticker is embedded into a facial image as the input to attack FR systems, which can obtain results similar to those of physical attacks without inviting any volunteers. The results show that IPAA has a higher similarity with physical attacks than digital attacks, indicating that IPAA is able to evaluate the performance of physical attacks. IPAA is effective in quantitatively measuring the impact of the sticker location on the results of attacks.

Keywords: physical attack; digital attack; face recognition; interferential variable; adversarial example
CLC number: TP391 **Document code**: A
Article ID: 1672-5220(2025)04-0442-07

Open Science Identity (OSID)

0 Introduction

Deep neural networks (DNNs) have shown great performance in many fields, such as face recognition (FR), object detection and natural language processing. However, DNNs are vulnerable to adversarial examples, which makes DNNs unsafe. A DNN-based system would give a wrong judgment with a carefully designed input[1-8]. For example, pasting an adversarial sticker onto the forehead causes the FR systems to fail to recognize the person[5]. Printing adversarial patterns on clothing prevents the object detectors from detecting pedestrians[9]. In information security, adversarial attacks play a beneficial role in protecting personal information by rendering unauthorized detectors ineffective[10-11].

Usually, adversarial attacks are classified into digital attacks and physical attacks. In the method of digital attacks, some pixels on the images are modified to generate the adversarial images, and they are fed directly into the DNN-based system to attack the DNN-based system. During the attack process, there is no distortion problem. Therefore, digital attacks always exhibit excellent attack performance in successfully deceiving the DNN-based system.

In the method of physical attacks, adversarial examples are created in the digital domain and subsequently transferred to the physical domain. In most cases, the transformation means printing out the adversarial examples, which may result in distortion due to the low printing resolution and color differences. This could potentially worsen the performance of physical attacks. When attacking the DNN-based system, the adversarial examples are transferred from the physical domain to the digital domain through photography. They are fed into the system and lead to incorrect output results. It can be seen that physical attacks have two more transformations than digital attacks. This results in the loss of some adversarial information, making the performance of physical attacks not as good as that of digital attacks.

In the complex physical environment, interferential variables such as lighting, background and photographing distance can cause the adversarial examples to distort and impair the efficacy of physical attacks[5, 9, 12]. Some interferential variables, such as the sticker location, are used to attack FR systems[13-14]. In order to improve the robustness of the adversarial examples, it is necessary to study the relationship between interferential variables and the results of physical attacks. However, it is a challenging task to measure how much the impact of a single interferential variable on the results of attacks without changing other interferential variables. In Ref. [15], adversarial infrared patches are used to prevent object detectors from detecting pedestrians. Researchers tested the results of attacks at six photographing angles and four photographing distances, but other interferential variables (e. g. human posture) might change during the testing process and affect the test results.

Calculating an attack success rate (ASR) requires

Received date: 2024-07-03
* Correspondence should be addressed to XU Wujun, email: wujun.hsu@qq.com
Citation: HU W T, XU W J. Incomplete physical adversarial attack on face recognition[J]. *Journal of Donghua University (English Edition)*, 2025, 42(4): 442-448.

many data samples to obtain reliable results. However, the ASR of physical attacks is complicated to obtain for researchers because of the lack of sufficient volunteers. In Ref. [16], only 10 people were invited as volunteers to test the ASR of physical attacks, and the error of the ASR is 10%. If they invited 100 people to test the ASR, the error of the ASR would be reduced to 1%. However, inviting 100 people would increase the workload and research cost for researchers.

To solve the above issues, an easy-to-implement attack method called incomplete physical adversarial attack (IPAA) is proposed. Different from the physical attack, IPAA requires taking photos of faces and adversarial examples separately, and then fusing the two images in the digital world to generate adversarial images. IPAA retains almost the same distortion information as the physical attack and is capable of effectively simulating the performance of the physical attack. By using IPAA, there is no need to invite any volunteers to test the results of the attacks, and just facial images are needed, thereby rendering IPAA a substitute for the physical attack. Furthermore, by taking photos of faces and the adversarial examples separately, researchers can flexibly control the interferential variables in the experiment. This enables the quantitative measurement of the relationship between interferential variables and the results of the attack.

1 Methods

1.1 Principle of IPAA

A sticker-based method is used to conduct targeted attacks on FR systems. The original image is used to train the adversarial sticker in the digital domain. The face and printed sticker are photographed separately, and the image of the sticker is embedded into the forehead of the facial image to generate the IPAA image. The relationship between the original image and IPAA image is shown in Fig. 1.

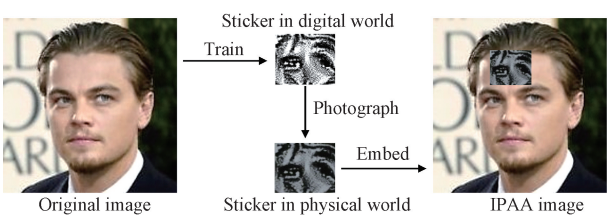

Fig. 1 Relationship between original image and IPAA image

The generative process of IPAA images is similar to that of physical attack images. The difference is shown in Fig. 2. The left area marked as D represents the digital domain, and the right area marked as P represents the physical domain. The symbol $\boldsymbol{F}_i$ ($i = 1, 2, 3$) represents different faces: $\boldsymbol{F}_1$ falls into the digital domain, indicating that it is not a real facial image (maybe a virtual facial image generated by AI); $\boldsymbol{F}_2$ falls into the physical domain, indicating that it is a real human face in the physical world; $\boldsymbol{F}_3$ falls into the intersection of D and P, indicating that it is a photo of a real human face. The symbol $\boldsymbol{S}_i$ ($i = 1, 2, 3$) represents different adversarial stickers: $\boldsymbol{S}_1$ is a digital adversarial sticker image in the digital domain; $\boldsymbol{S}_2$ is a real sticker printed from $\boldsymbol{S}_1$ in the physical domain; $\boldsymbol{S}_3$ is a photo of $\boldsymbol{S}_2$.

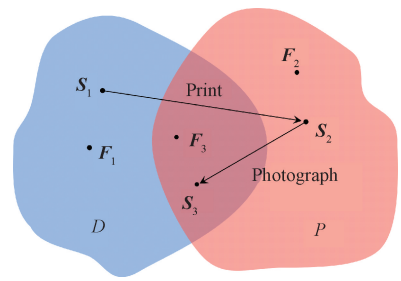

Fig. 2 Different faces and stickers in digital domain and physical domain

In the task of physical attacks, an adversarial image is generated by adding the perturbation $\boldsymbol{r}$ to a facial image $\boldsymbol{x}$ and printed as $\boldsymbol{S}_2$. Then $\boldsymbol{S}_2$ is stuck to $\boldsymbol{F}_2$ and photographed as the input of the FR systems. It should be noted that both the printing and photographing processes may introduce distortions in the adversarial stickers. This is due to the influence of interferential variables, including the resolution employed.

Suppose $\boldsymbol{\theta}$ represents the information of interferential variables. Then the adversarial image of physical attacks $\boldsymbol{x}^{\text{adv}}$ is

$$\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} + \boldsymbol{r} - \boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{r} - \boldsymbol{\theta}$ represents that the adversarial information is reduced by interferential variables. When it is a digital attack, $\boldsymbol{x}^{\text{adv}}$ equals $\boldsymbol{x} + \boldsymbol{r}$ without $\boldsymbol{\theta}$, as the digital attack is not affected by interferential variables. It can be assumed that the results of the digital attack represent the optimal outcome. The introduction of interferential variables will cause the results of physical attacks to be away from those of digital attacks. Therefore, the performance of physical attacks is likely to be inferior to that of digital attacks.

The attack process of IPAA also contains printing and photographing processes. It will produce distortion in the adversarial sticker. The similar distortion between the physical attack and IPAA will lead to their similar performance. IPAA is an incomplete physical attack because stickers are individually photographed and embedded into digital facial images, rather than being directly stuck onto real faces. Figure 3 shows the difference between the physical attack and IPAA in the generation of attack images. In Fig. 3 (a), the man with the adversarial sticker on his forehead is photographed to generate the physical attack image. In Fig. 3 (b), the man and the adversarial sticker are photographed separately, and the sticker image is embedded into the facial image to generate the IPAA image.

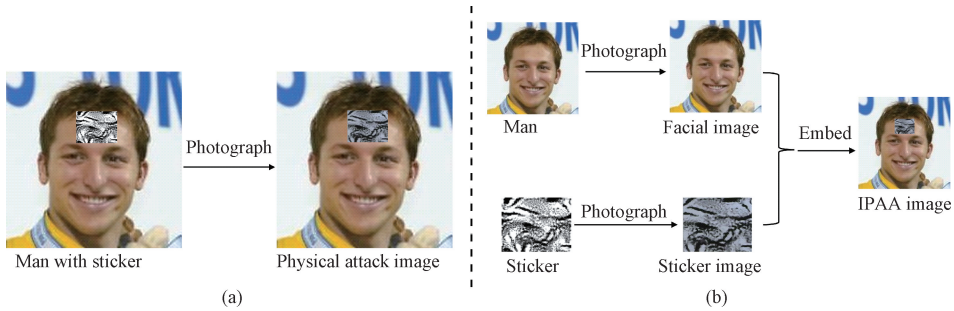

Fig. 3 Comparison of the generation process of different attack images: (a) physical attack image; (b) IPAA image

The similarity between the physical attack and IPAA is that both facial images and sticker images come from the physical world rather than the digital world. The images of faces and stickers contain the information about distortion from the interferential variables, which ensures that the performance of a sticker image embedded into a facial image is almost the same as that of a real sticker stuck on a real face. Considering that one facial image is sufficient for IPAA to generate adversarial images, it is unnecessary to invite any volunteers to generate adversarial images to test the ASR of physical attacks.

1.2 Measurement indicators

Usually, an ASR is used to evaluate the performance of physical attacks. However, the ASR is a statistical value that cannot describe the result of each attack, and it fluctuates in values when the threshold of the FR model is modified. If the threshold is set improperly or the recognition accuracy is not the highest, it may lead to a high or low ASR. Therefore, the ASR is not suitable to be the results of measurement.

Let $\boldsymbol{e}_{\mathrm{t}}$ be the output feature vector of the target facial image, and $\boldsymbol{e}_{\mathrm{a}}$ be the output feature vector of the adversarial facial image. The cosine distance between $\boldsymbol{e}_{\mathrm{t}}$ and $\boldsymbol{e}_{\mathrm{a}}$ is defined as

$$d = 1 - \frac{\boldsymbol{e}_{\mathrm{t}} \cdot \boldsymbol{e}_{\mathrm{a}}}{\|\boldsymbol{e}_{\mathrm{t}}\| \cdot \|\boldsymbol{e}_{\mathrm{a}}\|}, \tag{2}$$

where d represents the degree of similarity between the target image and the adversarial image without any statistical processing. It reflects the result of a single attack and is not affected by the threshold of the FR model, so it is used to calculate the results of the measurement. A lower cosine distance indicates a higher degree of similarity between two vectors.

1.3 Similarity between attack methods

A mean square error (MSE) is employed to calculate three types of differences in the cosine distance: the difference between the digital attack and physical attack (D&P), the difference between the physical attack and IPAA (P&I) and the difference between the digital attack and IPAA (D&I). According to the previous discussion in subsection 1.1, the difference between digital attacks and physical attacks is caused by the distortion from interferential variables. Therefore, the MSE represents the distortion distance between attack methods:

$$\psi(\boldsymbol{u}, \boldsymbol{v}) = \frac{1}{n}\sum_{i=1}^{n}(u_i - v_i)^2, \tag{3}$$

where $\psi(\boldsymbol{u}, \boldsymbol{v})$ is the distortion distance between the vector $\boldsymbol{u}$ and the vector $\boldsymbol{v}$, and $\boldsymbol{u}$ or $\boldsymbol{v}$ is the vector of the cosine distance for one of the three attack methods; n is the number of samples.

Digital attacks have the best attack performance, and are considered as the benchmark to calculate the similarity. The similarity is defined as

$$\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \frac{2\psi(\boldsymbol{u}, \boldsymbol{v})^{-1}}{\sum_{\boldsymbol{p}\neq\boldsymbol{q}}\psi(\boldsymbol{p}, \boldsymbol{q})^{-1}}, \tag{4}$$

where $\boldsymbol{p}$ or $\boldsymbol{q}$ is the vector of the cosine distance for one of the three attack methods; $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v})$ means the proportion of one distortion distance to the total distortion distances.

1.4 Generation of adversarial stickers

Grayscale images with one channel are used to create adversarial stickers instead of RGB images with three channels. Although grayscale images contain less adversarial information than RGB images, they are easier to print and have lower production costs compared with RGB images.

To embed the grayscale image of the sticker into the RGB image of the face, it is necessary to expand the single channel of the grayscale image to three channels. Furthermore, the gradient must be weighted to the three channels during the training process. The final gradient is

$$\mathrm{grad}(\boldsymbol{M}^{h\times w}) = \boldsymbol{C} \times \nabla \boldsymbol{M}^{3\times h\times w}, \tag{5}$$

where $\boldsymbol{C}$ is the weight coefficient vector of the three channels, and $\boldsymbol{C} = (c_1, c_2, c_3)$, $\sum_{i=1}^{3} c_i = 1$; $\boldsymbol{M}^{3\times h\times w}$ is the RGB image of the adversarial sticker with three channels; $\boldsymbol{M}^{h\times w}$ is the grayscale image of the adversarial sticker; $\mathrm{grad}(\boldsymbol{M}^{h\times w})$ is the gradient of the sum of the three channel weights; $\nabla \boldsymbol{M}^{3\times h\times w}$ is the gradient of $\boldsymbol{M}^{3\times h\times w}$.

The adversarial sticker is trained by using the I-FGSM[17] algorithm:

$$\boldsymbol{x}_k^{\mathrm{adv}} = \mathrm{clip}(\boldsymbol{x}_{k-1}^{\mathrm{adv}} + \alpha \cdot \mathrm{sign}(\nabla \boldsymbol{J}(\boldsymbol{x}_{k-1}^{\mathrm{adv}}))), \tag{6}$$

where $k \in \mathbf{N}$ is the number of iterations; $\boldsymbol{x}_k^{\mathrm{adv}}$ is the k_{th} adversarial image; clip $(\cdot)$ restricts the pixels of the

adversarial image within the valid range after perturbation; α is the perturbation step size; $\nabla \boldsymbol{J}(\boldsymbol{x}_{k-1}^{\text{adv}})$ is the gradient of $\boldsymbol{x}_{k-1}^{\text{adv}}$.

1.5 Quantitative measurement

There are many interferential variables that can worsen the performance of physical attacks. In order to quantitatively measure the impact of the single interferential variable on physical attacks, the other variables should remain unchanged. When using the IPAA method, after taking photos of the face and sticker separately (e. g. Fig. 3 (b)), some interferential variables, such as lighting, are fixed on the photos and remain unchanged. Conversely, other interferential variables, such as the sticker location, can be manipulated freely on the computer.

When embedding the sticker image into the facial image, the adversarial image $\boldsymbol{x}^{\text{adv}}$ is defined as

$$\boldsymbol{x}^{\text{adv}} = \boldsymbol{x} \odot (1 - \boldsymbol{A}) + \boldsymbol{M} \odot \boldsymbol{A}, \tag{7}$$

where $\boldsymbol{A}$ is the mask to determine the location of the adversarial sticker $\boldsymbol{M}$ on the facial image $\boldsymbol{x}$.

The multi-task convolutional neural network (MTCNN)[18] is a model for the face detection task. The five main points of the face are detected by MTCNN to calculate the location and scale of the mask on the facial images.

2 Experiments

Three experiments are designed in this section. The first one is to verify the similarity between attack methods. The second one is to utilize IPAA to measure the influence of the sticker location. The third one is to verify the defense against IPAA.

2.1 Verification of similarity between attack methods

The experiment does not focus on the performance of the adversarial algorithm but on the similarity between attack methods. The targeted attacks are conducted in the experiment to make the adversarial facial image be recognized as a specific individual by FR systems.

Four different FR models are used as target models, including VGGFace[19], FaceNet[20], ArcFace[21] and MobileFace[22]. To provide original facial images, 25 people are invited as volunteers, including 7 females and 18 males. Their ages are between 16 and 60 years old. Four facial images are selected randomly from the labeled faces in the wild (LFW)[23] dataset as the target facial images. Each original image will generate an adversarial example by using one of the target images.

During the training process, each adversarial example is generated after 200 iterations by using the I-FGSM algorithm. The perturbation step size α is set to 0.005. To verify the similarity between attack methods, a procedure is designed to measure the cosine distance between attack methods. The whole verification process is shown in Fig. 4, where d_{TD} represents the cosine distance between target facial images and digital attack images; d_{TP} represents the cosine distance between target facial images and physical attack images; d_{TI} represents the cosine distance between target facial images and IPAA images.

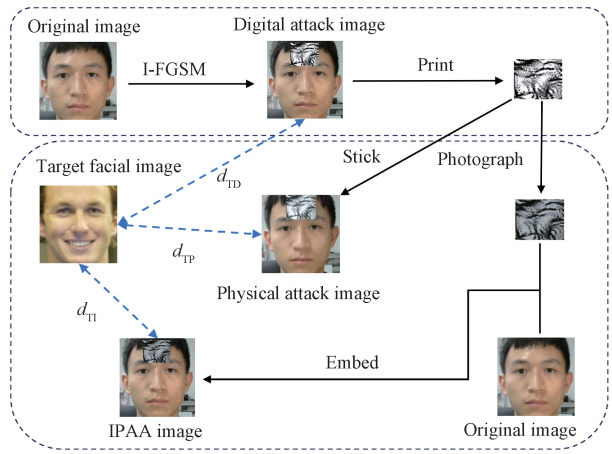

Fig. 4 Whole verification process of IPAA

The I-FGSM algorithm is used to train the digital attack image in the digital world. The adversarial sticker is extracted from the digital attack image and printed out as the adversarial sticker. Firstly, the sticker is photographed individually and embedded into the original image to generate the IPAA image. Then, the sticker is stuck onto the volunteer's forehead and photographed to generate the physical attack image. Finally, the digital attack image, physical attack image and IPAA image with the target facial image are input into the FR model, respectively, to calculate their cosine distance.

Take VGGFace as an example. The cosine distance curves of three attack methods are plotted in Fig. 5, where no attack refers to the cosine distance between the original images and target images. All the cosine distances of digital attacks are the smallest, indicating that digital attacks have the best attack performance. The cosine distances of the physical attack and IPAA are very close. This observation suggests that IPAA has a similar performance to the physical attack.

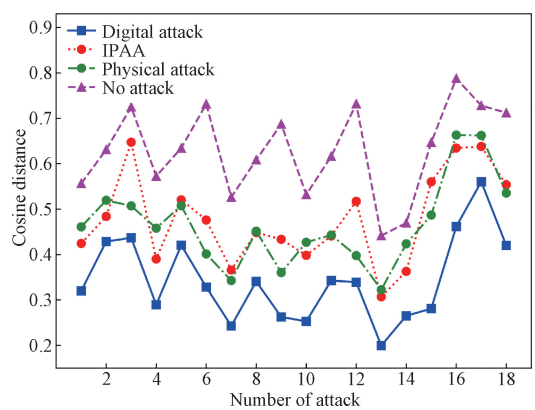

Fig. 5 Cosine distances of digital attack, physical attack, IPAA and no attack

The distortion distance and similarity between attack methods are calculated by Eqs. (3) and (4), and the results are listed in Tables 1 and 2. P&I has the smallest distortion distance. In Table 2, the similarity of P&I is the highest. In comparison to the digital attack, IPAA is closer to the physical attack. It can be concluded that IPAA can be used to simulate the physical attack to obtain the ASR of the physical attack within a small margin of error.

Table 1 Distortion distance between attack methods

Attack method	Distortion distance			
	VGGFace	FaceNet	ArcFace	MobileFace
D&I	0.020 7	0.045 7	0.010 2	0.034 7
D&P	0.016 5	0.069 4	0.023 4	0.081 2
P&I	**0.003 6**	**0.006 4**	**0.005 2**	**0.016 9**

Table 2 Similarity between attack methods

Attack method	Similarity/%			
	VGGFace	FaceNet	ArcFace	MobileFace
D&I	12.49	11.38	29.58	28.71
D&P	15.67	7.49	12.89	12.29
P&I	**71.84**	**81.13**	**57.53**	**59.00**

The distortion distance is the difference between the attack methods caused by interferential variables. Assume that there is a space consisting of distortion vectors from interferential variables, and the digital attacks with no distortion are placed at the original point *O*. The location of the three attack methods is shown in Fig. 6. The distortion distance of the physical attack or IPAA is generated by the superposition of distortion vectors.

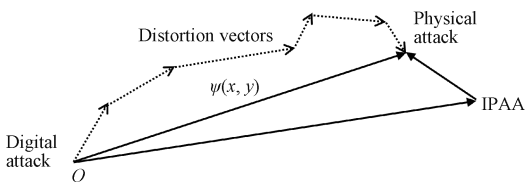

Fig. 6 Distortion distance of three attack methods

During the experiment, the distortion vectors of the physical attack and IPAA are not exactly the same, resulting in different distortion distances. Therefore, their similarity will not reach 100%.

The distortion distance between digital attacks and physical attacks is caused by interferential variables. If digital attacks gradually approach physical attacks, the ASR of physical attacks will also gradually approach that of digital attacks, thereby improving the performance of physical attacks.

2.2 Measurement of effect of sticker location

The following experiment is designed to obtain the relationship between the sticker location and results of attacks. MTCNN is used to automatically adjust the scale and location of the sticker image. When embedding the sticker image into the facial image, the sticker image is offset to the left, right, top and bottom separately with different pixel distances. The offset range is limited to 20 pixels. In this way, 400 IPAA images with different locations of the sticker on the facial image are generated. Then, these IPAA images are input into the VGGFace model to calculate the cosine distance between the target images and IPAA images.

The relationship between the sticker location and cosine distance is shown in Fig. 7, where x and y correspond to the coordinates of the sticker on the facial image. The optimal coordinate with the minimum cosine distance is located at the bottom of the graph. In these 400 times of attacks, the range of cosine distance is [0.411, 0.514], and the optimal coordinate is (75, 27). According to the previous experimental results, it is thought that the best result of physical attacks is 0.411 when the sticker is at (75, 27). Furthermore, the vertical offset of the sticker has a greater impact on the results of attacks than the horizontal offset. This indicates that the results of attacks are more sensitive to the vertical offset than the horizontal offset. To improve the robustness of adversarial examples against variations in their locations within an image, it is necessary to flatten the curved surface.

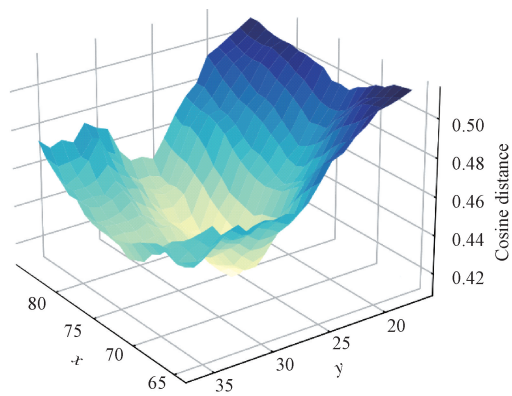

Fig. 7 Relationship between sticker location and cosine distance

The impact of the sticker location on the results of physical attacks can be quantitatively measured by IPAA. However, it is impossible to measure the relationship through physical attacks, as the experiment requires moving the sticker 400 times on the volunteer's forehead and each time ensuring that the interferential variables (e.g. facial expression) remain unchanged.

2.3 Defense against IPAA

Due to the high similarity between the physical attack and IPAA, the defense methods against the physical attack can also be applied to IPAA. In this experiment, the IPAA images are processed by Gaussian filtering and bilateral filtering to destroy the adversarial information before being input into the FR model. The output results are shown in Fig. 8. The cosine distances of the processed IPAA images are higher than those of the

unprocessed IPAA images. That means the processed IPAA images will not be recognized as the targeted images, and the defense method is effective against IPAA.

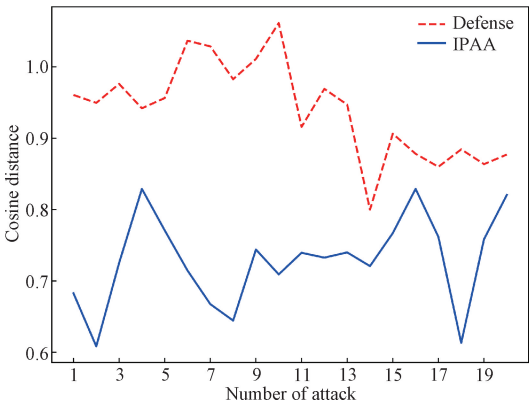

Fig. 8 Cosine distances of IPAA and defense

3 Conclusions

In this paper, an easy-to-implement attack method called IPAA is proposed to simulate the physical attack. Experiments conducted on four FR models demonstrate that the physical attack and IPAA exhibit a high degree of similarity (81.13%). IPAA can be used to simulate the physical attack within a certain margin of error. By using IPAA, there is no need to look for any volunteers but facial images to test the ASR of the physical attack. This significantly simplifies the research process. Additionally, IPAA is employed to quantitatively measure the relationship between the sticker location and results of the physical attack. The experiment of 400 attacks with different locations finds that the optimal attack result is 0.411.

References

[1] LIU X L, SHEN F R, ZHAO J, et al. EAP: an effective black-box impersonation adversarial patch attack method on face recognition in the physical world [J]. *Neurocomputing*, 2024, 580: 127517.

[2] KHEDR Y M, LIU X, HE K. TransMix: crafting highly transferable adversarial examples to evade face recognition models[J]. *Image and Vision Computing*, 2024, 146: 105022.

[3] HU C, LI Y B, FENG Z H, et al. Attention-guided evolutionary attack with elastic-net regularization on face recognition [J]. *Pattern Recognition*, 2023, 143: 109760.

[4] AGRAWAL K, BHATNAGAR C. A black-box based attack generation approach to create the transferable patch attack [C]//2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS). New York: IEEE, 2023: 1376-1380.

[5] ZHENG X, FAN Y B, WU B Y, et al. Robust physical-world attacks on face recognition [J]. *Pattern Recognition*, 2023, 133: 109009.

[6] SURYANTO N, KIM Y, KANG H, et al. DTA: physical camouflage attacks using differentiable transformation network [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 15284-15293.

[7] SCHNEIDER J, APRUZZESE G. Dual adversarial attacks: fooling humans and classifiers [J]. *Journal of Information Security and Applications*, 2023, 75: 103502.

[8] HU C Y, SHI W W, TIAN L. Adversarial color projection: a projector-based physical-world attack to DNNs [J]. *Image and Vision Computing*, 2023, 140: 104861.

[9] HU Z H, HUANG S Y, ZHU X P, et al. Adversarial texture for fooling person detectors in the physical world [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2022: 13297-13306.

[10] ZHAO M N, WANG B, GUO W K, et al. Protecting by attacking: a personal information protecting method with cross-modal adversarial examples [J]. *Neurocomputing*, 2023, 551: 126481.

[11] QIN Y X, ZHANG K J, PAN H W. Adversarial attack for object detectors under complex conditions [J]. *Computers & Security*, 2023, 134: 103460.

[12] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples [C]//International Conference on Machine Learning. Stockholm, Sweden: IMLS, 2018: 284-293.

[13] RYU G, PARK H, CHOI D. Adversarial attacks by attaching noise markers on the face against deep face recognition[J]. *Journal of Information Security and Applications*, 2021, 60: 102874.

[14] WEI X X, GUO Y, YU J. Adversarial sticker: a stealthy attack method in the physical world[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 2711-2725.

[15] WEI X X, YU J, HUANG Y. Physically adversarial infrared patches with learnable shapes and locations[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2023: 12334-12342.

[16] KOMKOV S, PETIUSHKO A. AdvHat: real-world adversarial attack on ArcFace face ID system[C]//2020 25th International Conference on Pattern Recognition (ICPR). New York: IEEE, 2021: 819-826.

[17] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[EB/OL]. (2016-07-08)[2024-06-20]. http://arxiv.org/pdf/1607.02533.
[18] ZHANG K P, ZHANG Z P, LI Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499-1503.
[19] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep face recognition[C]//Proceedings of the British Machine Vision Conference 2015. Swansea: British Machine Vision Association, 2015: 41.1-41.12.
[20] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York: IEEE, 2015: 815-823.
[21] DENG J K, GUO J, YANG J, et al. ArcFace: additive angular margin loss for deep face recognition[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, 44: 5962-5979.
[22] CHEN S, LIU Y, GAO X, et al. MobileFaceNets: efficient CNNs for accurate real-time face verification on mobile devices [C]//13th Chinese Conference. [S. l.]: CCBR, 2018.
[23] Computer Vision Lab. Labeled faces in the wild [D]. Massachusetts: University of Massachusetts Amherst, 2007.

针对人脸识别的不完全物理对抗性攻击

胡伟涛[1], 许武军[1,2*]
1. 东华大学 信息科学与技术学院, 上海 201620
2. 东华大学 数字化纺织服装技术教育部工程研究中心, 上海 201620

摘　要: 在最近的研究中, 对抗性贴纸被广泛用于攻击物理世界中的人脸识别 (face recognition, FR) 系统。然而, 由于在实验中缺乏志愿者, 很难评估物理攻击的性能。该文提出了一种简单的攻击方法, 称为不完全物理对抗性攻击 (incomplete physical adversarial attack, IPAA), 用来模拟物理攻击。与物理攻击的过程不同的是, 在发起 IPAA 时, 对抗性贴纸的照片会嵌入人脸图像中, 作为输入图像来攻击 FR 系统, 从而在不邀请任何志愿者的情况下获得近似于物理攻击的结果。结果显示, IPAA 与物理攻击的相似性高于数字攻击, IPAA 能够评估物理攻击的性能。IPAA 也可以有效地定量测量贴纸位置对攻击结果的影响。

关键词: 物理攻击; 数字攻击; 人脸识别; 干扰变量; 对抗样本