A white-box impersonation attack on the FaceID system in the real world

Lu Guo¹, Hua Zhang*

1 The State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China¹,*
18810167207@163.com¹, zhanghua_288@bupt.edu.cn*

Abstract. The arcface model maximizes the classification boundary in the angular space, and is one type of the best models in the current face recognition model. We propose a mask sticker attack method to realize the impersonation attack of arcface model. The method specifically uses a parabolic transformation to simulate the bending situation of a sticker on a mask, and uses a multi-stage PGD attack to generate an adversarial sticker. Finally, the adversarial sticker is attached to the mask worn by the attacker to perform an impersonation attack. On targets of different skin colors, ages and genders, the method proposed in this article has an attack success rate of 0.65, and the final cosine similarity between attacker with mask sticker and target can reach about 0.5-0.7. In addition, we discuss the effect of sticker location and size on attack effect, as well as the generalization of this method on other models.

1. Introduction
In recent years, due to the continuous improvement of computing power and the continuous emergence of large data sets, deep learning methods have shown good performance. In the field of face recognition, the performance of verification [1] and recognition [2] is even better than humans. The recently proposed arcface model[3] is an improvement of the previous face recognition models, which use the loss of the angle space to replace the loss of previous cosine space in cosface model[4], and even the loss of the earlier Euclidean distance space in FaceNet model[5]. The latest loss achieves higher similarity of all face image of the same person, and lower similarity among different people, which has been verified in comparative experiments by Deng et al.[3] In addition, in some face recognition competitions such as the Megaface competition, the arcface solution is comparable to the models of Microsoft and Google. There are many open source data sets such as LFW[6], CASIA-WebFace[7] for researchers to use. Therefore, researchers can train an enterprise-level face recognition model by themselves.

Some studies have proved that above face recognition models based on deep learning are very susceptible to the interference of small disturbances. There are many attack methods in the electronic world, such as the decision-based black box attack algorithm proposed by Dong et al. [8], which realized the attack on the arcface model; Hakon et al. [9] proposed a new conditional generative adversarial network model, DeepPrivacy, which performed face-changing operation to achieve the attack. Above attacks in the electronic world mainly modify the picture pixels and cannot work in the real world.

At present, there have been many attacks on face recognition systems in the real world. Feng et al. [10] attacked Eigenfaces [11] by makeup and changing hairstyle. Yamada et al. [12] used the camera's...
sensitivity to near-infrared light to design a kind of light-emitting glasses, which can avoid the
detection of face recognition system. Erdogmus et al. [13] used 3D printed masks or pictures
downloaded from the Internet to attack. Mahmood et al.[14] pointed out that this method is easily
blocked by anti-fraud measures, and proposed a method using gradient descent to generate glasses
frame stickers to attack Eigenfaces [11], and the success rate of attack can reach 0.75 in most cases. In
order to achieve the universal attack by using eyeglass frame sticker, Mahmood et al. proposed AGN
(Adversarial Generative Nets)[15] based on GAN, which generates an eyeglass frame sticker similar to
normal glasses to attack VGG model, and this attack method had an attack success rate of more than
0.67 in probability of 0.5. Stepan et al. [16] recently designed a method of putting stickers on hats to
dodge the arcface model, which first revealed the arcface model based on deep learning. They could
not achieve impersonation attack.

By studying the previous attacks on the face recognition system in the real world, we found that all
the methods have not tried the impersonation attack of the open source face recognition system with
the best effect at this stage, the cost of some attack methods is relatively high, such as makeup[10],
luminous glasses[12], and some attack methods such as 3D printed masks or pictures downloaded
online[13] are easily blocked by anti-fraud methods such as life detection.

Contribution. This paper implements an impersonation attack on arcface model, which is the most
advanced face recognition model in real world. Our attack method consumes lower cost and is not
easy to be detected by anti-fraud methods. The contributions are follows.

● In this article, we propose an adversarial sticker generation algorithm based on multiple iterations
and one-step descent method, which can be used in reality.

● Based on a large number of attack experiments on LResnet100e IR model, we found that the
attack effect of the sticker in the middle of the mask is better than that in other positions, and the larger
the size of the sticker, the better the attack effect. The success rate can reach about 0.65, and the final
cosine similarity between attacker with mask sticker and target can reach about 0.5-0.7.

The following part of this paper is composed of the following chapters. The second section
introduces the work related to adversarial attack in the electronic world and the physical world. The
third section proposes a mask sticker attack against the face verification system. In the fourth section,
the effectiveness and the generalization of the proposed attack method is analyzed from the size,
location, skin color, age and gender of the target. The fifth part summarizes the work of this paper and
puts forward the future research direction.

2. Related work
In recent years, people have paid more and more attention to the security issues of deep models.
Szegedy et al.[17] first proposed the concept of adversarial examples in 2013. Adversarial example
attacks mainly refer to adding some disturbances to the picture, making the output of the neural
network wrong. The attack that greatly reduces the confidence of the original class of output and
greatly increases the confidence of any other class is called no target attack, also known as escape
attack. The attack with the highest confidence of a specific type of output is targeted attack, also
known as impersonation attack.

Attacks in the electronic world are mostly imperceptible to the human eye. According to the range
of modified pixels, they can be roughly divided into two categories. One is to modify the pixels in the
whole picture range. This type of attacks include FGSM [18], which is one-step attack based on
gradient descent and the I-FGSM method [19] based on multiple iterative gradient descent, the
depfool method[20] that crosses the decision boundary through multiple iterations, and the PGD
method [21] ,which is the strongest step-by-step attack . Another type of attacks is to modify part of
the image pixels. This kind of attack includes JSMA method [22] using forward derivative, which
mainly attacks by modifying the pixels that have great influence on the target model, and the one pixel
method[23] which attacks by modifying only one pixel.

Most of the attacks that are easily detectable by the human eye are in the form of stickers and
posters. These attacks are generally used in real world. Due to the complicated real-world environment,
it was not until Athalye et al. proposed the EOT method [24] that it was truly feasible to attack deep learning models in the real world. This method mainly simulates the complex environment in the real world, such as different illumination, angle transformation, lens noise, and so on. In addition, NPS and TV loss are proposed in the Against Glasses Frame Attack [15]. These two losses make the attack effect better in the real world. Because some colors in the RGB image displayed on the computer may not be printed by the printer, it will have an impact on the attack effect. The emergence of NPS mainly solves this problem by making the generated pixels closer to the printable color to reduce the loss caused by the printing process. TV loss is to make the generated stickers smoother.

Because of the good effect of the above three methods, many subsequent papers have used these methods. For example, EOT was first used in the adversarial patch method[25]mainly aimed at image classification models; in the paper[26], EOT and NPS are used to make confrontation posters or stickers to attack the traffic signal recognition model; the paper[27] also uses EOT to attack the traffic signal recognition model; a small part of the work[28,29] uses EOT, NPS, TV loss to attack the faster R-CNN, YOLO V2 model[30] too.

Above methods do not include all realistic scenes. They can achieve good attack effects in the case of pasting stickers on flat surfaces, but cannot guarantee the attack effect in the case of sticking stickers on curved surfaces, such as pasting stickers on real faces. The paper[16] proposed a method that can put stickers on real faces by Parabola deformation in 3D. This method can achieve escape attacks on face verification systems in the real world, but can not achieve impersonation attacks.

3. Mask stickers impersonate attacks

3.1. Motivation
According to the prior knowledge, the face recognition model is sensitive to the features of the key parts of the face, such as eyes, nose and mouth. Therefore, we try to change the key parts of the face to achieve an easier and better attack. As some scholars are studying the attack on the eye area and have achieved some results, we would like to try attacks on other key areas. Due to the impact of the new crown epidemic, more and more people need to wear masks every day, and even some face recognition systems of BUPT (Beijing University of Posts and Telecommunications) can detect the wearing of masks. Therefore, we focus our research work on the mouth and nose, and try to generate the adversarial stickers suitable for masks.

3.2. Workflow of mask sticker attack
We first convert the flat sticker into a curved sticker and place it on the mask position of the attacker's face picture, and adjust the combined picture size to the input size (112*112) required by arcface, then input the picture into the LResNet100E-IR model to obtain the 512-dimensional vector $e_h$ by embedding layer. The face image of the attack target is adjusted to 112*112, and the vector $e_t$ is obtained by inputting it into the same model. Calculate the adversarial loss $L_{adv}$ through $e_h$ and $e_t$, and add NPS, TV loss, then use the three-stage PGD attack method[21] to obtain the final sticker picture. The process is shown in the figure1.
3.3. Algorithm description

3.3.1. Bending simulation and mapping transformation of the sticker on mask. First of all, we simulate the deformation of the sticker on the mask in the real world. Since the human face is not a plane, and the sticker pasted on the mask is curved, we need to bend the flat sticker first, and map it to the mask of the plane face image through projection transformation. We use the parabola transformation method in 3D proposed in paper[16]. The sticker is on the mask and the camera is generally parallel to the face, so we do not use the pitch angle transformation after parabola transformation.

After parabolic transformation of the sticker, we also use the STN model[31] to map the curved sticker to the mask position of the flat face image to simulate the face of the mask with sticker. Since the input of the arcface model must be 112*112, we adjust the face image of the mask with stickers to the required size.

3.3.2. Loss. In this paper, our attack system is FaceID, which is 1:1. This system determines whether the two are the same person, which is only related to the two and has nothing to do with other people. Therefore, when we construct loss, we only need to consider the similarity between the attacker with mask sticker and the target. Since there is more than one face attacking the target in face recognition system, we consider the average similarity between the target's multiple face images and the attacker as the final adversarial loss as shown as formula (1).

$$L_{adv} = \frac{1}{n} \sum_{i=1}^{n} (1 - \cos(e_a, e_{ti}))$$

(1)

Where $n$ represents the number of face images of an attack target, $e_a$ refers to the feature of attacker through the arcface model, $e_{ti}$ represents the features of i-th attack target face image through the arcface model. $\cos()$ is the cosine distance.

In order to reduce the impact of printing and make the image smoother and attack more stable on devices with different image difference methods, we use the NPS and TV loss proposed in paper[15]. The NPS of pixel $p'$ is shown as formula (2) and TV loss as shown as formula (3).

$$NPS(p') = \prod_{p \in P} |p' - p|$$

(2)

Where $P \subset [0,1]^3$ is the set of printable RGB, if $p' \in P$ or $p'$ is close to a certain $p (p \in P)$, then $NPS(p')$ will be very low.


\[
TV(x) = \sum_{i,j} \left( (x_{i,j} - x_{i+1,j})^2 - (x_{i,j} - x_{i,j+1})^2 \right) \frac{1}{2}
\]  

(3)

We get the final loss \( L_{\text{final}} \) by considering adversarial loss, NPS and TV loss, which is shown as formula (4).

\[
L_{\text{final}} = L_{\text{adv}} + \lambda_1 \cdot TV(x) + \lambda_2 \cdot NPS(x)
\]  

(4)

Where \( x \) is a sticker, \( \lambda_1 \) and \( \lambda_2 \) are hyperparameters.

4. Attack effect evaluation

We mainly study the influence of the position of the sticker on the mask and the size of the sticker on the attack effect, as well as the difficulty and effect of this method on the target with different skin color, age and gender. At the same time, we evaluate generalization of this method by attacking other models with sticker attackers.

In the whole experiment, the face image is set to 600*600, and attack target images we use are selected from CASIA-WebFace[7] and LFW[6], which is a test set for building an arcface model, and each person randomly selected three photos. In the experiment of testing the influence of size and location on attack effect, we choose black men(figure 2) as the target, which is quite different from the category of attacker (myself), so we can roughly test the attack effect. The generated sticker is shown in Figure 3.

The attack method used in the whole experiment is PGD attack[21] based on gradient and iteration, which is the strongest one step attack method at present. The attack process is mainly divided into three stages, and different iterative steps and learning rates are used in different stages.

![Figure 2 black men](image1)

![Figure 3 The sticker generated by figure 2](image2)

It can be seen from Figure 3 that the generated sticker is somewhat similar to the features of the target face.

4.1. The impact of different sizes on attack effect

We studied the attack effect in different widths and heights. The attack results are shown in the figure 4-6.

Then, the ordinates in all the images with coordinates indicate the similarity between the attacker with the adversarial sticker and the attacking target, and the abscissa represents the total number of iterations required.
It can be seen from Figure 4-6 that when the height of sticker is 200, the similarity between the attacker and the target is less than 0.5 regardless of the width, and the impersonation attack cannot be realized. When the height of sticker is 300, the similarity between the attacker and the target is between 0.5-0.6 (including the boundary). When it increases to 400, the similarity is between 0.55-0.67. The latter two can achieve impersonation attacks.

And at a fixed height, when reducing the sticker width, it can be seen that the similarity gradually decreases from the red line to the green line to the orange line to the blue line in the figure. It can be seen that no matter increasing the width or height of the sticker, the attack effect will be improved.

It can be seen from the above three pictures that the number of iterations required by the green line is less than that of other color lines in most cases. It can be seen that when the width is 800, the attack time is the least.

So we combined the time and the effect of imitating the attack, and finally chose 400*800 as the optimal strategy.

4.2. The impact of different positions on attack effect
Since the area occupied by the sticker on the mask is not very small, we only discussed the influence of the left, middle and right positions. Since the optimal size (400*800) is a bit large, we consider a slightly smaller size sticker (400*700). The experimental results are shown in the figure 7.
We can see that the attack effect of the middle position is better than that of the left and right sides. The main reason is that the sticker in the middle will block the nose and mouth, while the stickers on the left and right sides may not or only partially block the two key parts mentioned above. Therefore, the stickers generated by the latter have little impact on the key parts of the face, leading to poor attack effect.

**4.3. The degree of difficulty in attacking different skin color, gender and age groups**

We mainly discuss the difficulty and effect of attacking targets in various situations, which include black, white and yellow people, children, adults (18-30), middle-aged (31-50), old (after 60) and male and female.

Where w stands for white people, b stands for black people, y stands for yellow people, f stands for women, m stands for men, c stands for children, y stands for adults, y2 stands for mature age, and o stands for old age.

The generated part of the stickers are shown in Figure 8-9:
The two pictures above show the stickers generated by white people as targets, which are still similar to facial features. Probably because this is a recognition task in the face scene, the data sets are all faces, only the features of the face can produce attack effect.

The experimental results are shown in the figure 10-15:

Figure 9 The sticker generated by wfy2

Figure 10 The similarity when target is white women of different ages

Figure 11 The similarity when target is white men of different ages

Figure 12 The similarity when target is black women of different ages

Figure 13 The similarity when target is black men of different ages
It can be seen from Figure 10-15 that the total success rate of impersonation attacks can be as high as 0.65. The success rate of attacks on white races can reach 0.5, and the success rate of attacks on black and yellow races can reach 0.875. This success rate will fluctuate for different attackers and targets. Therefore, I personally attack yellow people and black people more easily than white people. However, other attackers may be different from me.

And from these figures, we can know that the attack success rate for female targets is 0.66, and the attack success rate for male targets is 0.83. This success rate will fluctuate for different attackers and targets. Therefore, it is easier for me to attack male targets than female targets. Similarly, other people as attackers may be different from mine.

It can be seen from the above six figures that there is no rule to follow for the attack effect on different age groups under the same skin color and gender, so it is impossible to judge which age group is more likely to attack.

4.5. Generalization of attack methods
We randomly select 10 images from the above data set of different gender, skin color and age as the test. The attack success rate can reach 0.3 on the LResNet50E model and 0.2 on the MobileFace model. It can be seen that the success rate of attack on other models has decreased, but the cosine similarity between attacker and target can be slightly improved.

5. Conclusions
In this paper, we propose a mask sticker impersonation attack method for the face recognition system. On data sets of different ages, genders, and skin colors, the attack success rate is as high as 0.65, and the cosine similarity is about 0.5-0.7. Through experiments, we know that the larger the sticker, the more centered, the better the attack effect. However, the generalization of the method proposed in this paper is not very strong, and the attack success rate can only reach 0.2-0.3 on other models. In addition, the generated stickers are similar to the facial features, so it is easy to be noticed. These above issues are our main follow-up research work.

References
[1] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
[2] Taigman, Yaniv, et al. "Web-scale training for face identification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[3] Jiankang Deng, Jia Guo, et al. "Arcface: Additive angular margin loss for deep face recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[4] Hao Wang, Yitong Wang, et al. "Cosface: Large margin cosine loss for deep face recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[5] Schroff, Florian, et al. "Facenet: A unified embedding for face recognition and clustering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[6] Tianyue Zheng, Weihong Deng, et al. "Cross-agelfw: A database for studying cross-age face recognition in unconstrained environments." arXiv preprint arXiv:1708.08197 (2017).
[7] Yi Dong, Lei Zhen, et al. "Learning face representation from scratch." arXiv preprint arXiv:1411.7923 (2014).
[8] Yinpeng Dong, Hang Su, et al. "Efficient decision-based black-box adversarial attacks on face recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[9] Hukkelås, Håkon, et al. "Deepprivacy: A generative adversarial network for face anonymization." International Symposium on Visual Computing. Springer, Cham, 2019.
[10] Ranran Feng, Balakrishnan Prabhakaran, et al. "Facilitating fashion camouflage art." Proceedings of the 21st ACM international conference on Multimedia. 2013.
[11] Peter N Belhumeur, Joo P Hespanha, et al. "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection." IEEE Transactions on pattern analysis and machine intelligence 19.7 (1997): 711-720.
[12] Takayuki Yamada, Seiichi Gohshi, et al. "Privacy visor: Method for preventing face image detection by using differences in human and device sensitivity." IFIP International Conference on Communications and Multimedia Security. Springer, Berlin, Heidelberg, 2013.
[13] Nesli Erdogmus and Sebastien Marcel. “Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect”. In Proc. IEEE BTAS, 2013.
[14] Mahmood Sharif, Sruti Bhagavatula, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 acm sigsac conference on computer and communications security. 2016.
[15] Mahmood Sharif, Sruti Bhagavatula, et al. "Adversarial generative nets: Neural network attacks on state-of-the-art face recognition." arXiv preprint arXiv:1801.00349 2.3 (2017).
[16] Komkov Stepan, and Aleksandr Petiushko. "Advhat: Real-world adversarial attack on arcface face id system." arXiv preprint arXiv:1908.08705 (2019).
[17] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[18] Lan Goodfellow, Jonathon Shlens, et al. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
[19] Kurakin, Alexey, Ian Goodfellow, et al. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).
[20] Moosavi-Dezfooli, Seyed-Mohsen, et al. "Deepfool: a simple and accurate method to fool deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[21] Madry Aleksander, Makelov Aleksandar, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
[22] Nicolas Papernot, Patrick McDaniel, et al. "The impersonations of deep learning in adversarial settings." 2016 IEEE European symposium on security and privacy (EuroS&P). IEEE, 2016.
[23] Jiawei Su, Danilo Vasconcellos Vargas, et al. "One pixel attack for fooling deep neural networks." IEEE Transactions on Evolutionary Computation 23.5 (2019): 828-841.
[24] Athalye, Anish, et al. "Synthesizing robust adversarial examples." International conference on machine learning. 2018.
[25] Tom B. Brown, Dandelion Mané, et al. "Adversarial patch." arXiv preprint arXiv:1712.09665 (2017).
[26] Eykholt, Kevin, et al. "Robust physical-world attacks on deep learning visual
classification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[27] Sitawarin, Chawin, et al. "Rogue signs: Deceiving traffic sign recognition with malicious ads and logos." arXiv preprint arXiv:1801.02780 (2018).
[28] Shang-Tse Chen, Cory Cornelius, et al. "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2018.
[29] Kevin Eykholt, Ivan Evtimov, et al. "Physical adversarial examples for object detectors." 12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18). 2018.
[30] Redmon, Joseph, et al. "YOLO9000: better, faster, stronger." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[31] Jaderberg Max, Karen Simonyan, et al. "Spatial transformer networks." Advances in neural information processing systems. 2015.