Robust Backdoor Attacks against Deep Neural Networks in Real Physical World

Mingfu Xue[1], Can He[1], Shichang Sun[1], Jian Wang[1], and Weiqiang Liu[2]
[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]College of Electronic and Information Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China
{mingfu.xue, hecan, sunshichang, wangjian, liuweiqiang}@nuaa.edu.cn

***Abstract*—Deep neural networks (DNN) have been widely deployed in various practical applications. However, many researches indicated that DNN is vulnerable to backdoor attacks. The attacker can create a hidden backdoor in target DNN model, and trigger the malicious behaviors by submitting specific backdoor instance. However, almost all the existing backdoor works focused on the digital domain, while few studies investigate the backdoor attacks in real physical world. Restricted to a variety of physical constrains, the performance of backdoor attacks in the real world will be severely degraded. In this paper, we propose a robust physical backdoor attack method, PTB (physical transformations for backdoors), to implement the backdoor attacks against deep learning models in the physical world. Specifically, in the training phase, we perform a series of physical transformations on these injected backdoor instances at each round of model training, so as to simulate various transformations that a backdoor may experience in real world, thus improves its physical robustness. Experimental results on the state-of-the-art face recognition model show that, compared with the methods that without PTB, the proposed attack method can significantly improve the performance of backdoor attacks in real physical world. Under various complex physical conditions, by injecting only a very small ratio (0.5%) of backdoor instances, the success rate of physical backdoor attacks with the PTB method on VGGFace is 82%, while the attack success rate of backdoor attacks without the proposed PTB method is lower than 11%. Meanwhile, the normal performance of target DNN model has not been affected. This paper is the first work on the robustness of physical backdoor attacks, and is hopeful for providing guideline for the subsequent physical backdoor works.**

***Index Terms*—Artificial intelligence security, Physical backdoor attacks, Deep neural networks, Face recognition, Physical transformations.**

I. Introduction

Deep neural networks (DNN) have achieved remarkable performance on various tasks in real physical world, such as face recognition [1], object detection [2], [3] and self-driving cars [4], etc. However, recent studies indicate that, the malicious attackers can embed backdoors into the DNN models [5], [6], [7], [8]. The attacked DNN model behaves normally on benign inputs, but when a specific backdoor instance is input, the model will perform the malicious behaviors that specified by the attackers, e.g., classifying the backdoor instance as the target class [5]. This type of attack against the deep learning models is known as the *backdoor attack*. To date, massive researches have been conducted on backdoor attacks. The methods of implementing the backdoor attacks against the DNN models can be divided into two categories: 1) directly modify the parameters or weights of the target DNN model to embed the backdoor [9], [10], [11]; 2) inject a small batch of well-designed backdoor instances into the training set to train the DNN, so as to embed the backdoor [5], [6], [12].

However, almost all of the existing backdoor attacks are conducted in the digital domain, while few studies have been studied in real physical world. To the best of the authors' knowledge, there is only one unpublished work [13] (a preprint in arXiv) explores the physical backdoor attacks. The authors collects 9 different triggers (e.g., headbands, earrings, etc.) in the real physical world, and take photos for the attackers wearing these accessories. These captured backdoor photos are submitted to the target face recognition model to launch the attacks. However, this work only investigated the backdoor attacks under the ideal physical conditions, where the attackers are facing the camera in a proper distance. In addition, the backdoor images used to launch the attacks are extremely similar to these injected backdoor instances in the training set of target DNN model, which can not satisfies the complex scenarios in real world attacks.

Compared to the backdoor attacks in the digital domain, the physical backdoor attacks are more difficult and more challenging. Different from the digital backdoor attacks which directly submits the images to the DNN model, the inputs of DNN model in real physical world are captured by the external cameras and processed (such as cropping and aligning) by the system. Restricted by various physical constraints (e.g., lighting, distance, angle, etc.), the backdoors in real world may fail to trigger the attacks, or the attack success rate is severely degraded. The common physical constraints are as follows. First, due to the rotation and angle variation of target object, the trigger that captured by the camera will be completely different. Second, under different lighting conditions and distances, the trigger in a backdoor instance that captured by the camera are different. Finally, environmental noises are introduced in the process of capturing and processing the photos. All the above factors will greatly constrain the effectiveness of backdoor attacks in the real physical world.

In this paper, we propose PTB (physical transformations for backdoors), a robust backdoor attack method in real physical world. The proposed PTB method performs a series of transformations on the injected backdoor instances, which simulates these physical transformations that a backdoor trigger may experience in real world, so as to improve its robustness in

the physical world. Specifically, at each iteration of model training, we perform five different transformations on the injected backdoor instance, including rotation, angle, distance, Gaussian noise and brightness transformation, which model the following physical constraints: 1) different rotations of backdoor trigger; 2) facing the camera with different angles; 3) launching the attack at different distances; 4) noises introduced by image capturing and preprocessing; 5) changes of lighting conditions. The key idea behind the proposed method is that, let the backdoor trigger experience these transformations within the procedure of model training. As a result, in the real physical world, the robustness and effectiveness of backdoor attacks can be maintained even the trigger undergoes these complex physical transformations.

The contributions of this paper are threefolds:

- For the first time, we explore the robust backdoor attacks in the real physical world. By modeling the distribution of transformations that a trigger may experienced in real world, the proposed method performs various physical transformations on these injected backdoor instances, which greatly improves the physical robustness of the backdoor trigger.
- We extend the backdoor attack in the digital domain to the real physical world, i.e., from the 2D plane to the 3D space. The proposed PTB method successfully address the influences of various physical constraints, which ensure the high attack performance of backdoor attacks under those complex physical conditions.
- We launch the practical backdoor attacks on the DNN based face recognition model (VGGFace [14]) on a large and realistic dataset (YouTube Aligned Face dataset [15]). Experimental results show that, by injecting only a very small ratio (0.5%) of backdoor instances, a high attack success rate can be achieved under complex physical conditions. Under complex physical conditions, the proposed PTB method can achieve the attack success rate of 82%, while the attack success rate of backdoor attacks without the proposed PTB method is lower than 11%. Moreover, under simple physical conditions, the attack success rate of the proposed PTB method is up to 100%, which is also higher than that without PTB method. Meantime, the normal performance of the target model has not been affected, which indicates that, the proposed backdoor attack method is concealed and the backdoor attack is difficult to be detected.

This paper is organized as follows. The related work is reviewed in Section II. The proposed robust physical backdoor attack method is elaborated in Section III. Experimental results and analysis are presented in Section IV. This paper is concluded in Section V.

II. Related Work

In this section, we review the related works on backdoor attacks, including the existing backdoor attacks in the digital domain, the vulnerability of backdoors, and backdoor attacks in the real physical world.

Backdoor attacks in the digital domain. A large number of researches have been conducted on the backdoor attacks in the digital domain. To date, there are two different strategies to implement the backdoor attacks: 1) modifying the internal network structure or weight of the target DNN model; 2) through data poisoning.

For the first attack strategy, the attacker is assumed to have the perfect knowledge of the target DNN model. In this way, he can directly insert the neuron-level backdoors into the target DNN to modify the structure [16], or maximize the activation of a specific neuron to construct the backdoor [11]. Besides, the attacker can also add the well-designed perturbations into the weight of a specific layer of the target DNN model to embed the backdoor [17], [18], or flip the bits of weight values to inject the backdoor [9].

For the second attack strategy, the attacker does not requires to know the knowledge of the target DNN model. He only needs to inject a small batch of backdoor instances into the clean training set to train the target model, thus such attack strategy is more feasible for real-world attacks. Gu et al. [5] proposes BadNets, which pastes the specific signs (the yellow square sticker and the flower) onto the clean images to generate the backdoor instances. However, the backdoor triggers used in this work are obvious, thus can be noticed by humans. Therefore, a series of works have been performed to improve the concealment of the backdoors. For example, the attacker can add the adversarial perturbations into a clean image as the backdoor [12], or using the steganography technique to hide the backdoor in an image [19]. Recently, Liu et al. [20] indicate that, the reflections on the surface of smooth objects (such as glass) can also be used as the invisible backdoor to trigger the attack.

Vulnerability of backdoors. For backdoor attacks, the backdoor used to trigger the backdoor attack in test time should be consistent with the one that injected in the training phase. However, in real-world attacks, such condition may not be satisfied due to the processing operations of DNN model or the physical constraints, which will greatly degrade the effectiveness of backdoor attacks. Li et al. [21] explore the the characteristics of the backdoors, and demonstrate that if the location or shape of the trigger has been slightly changed, the performance of backdoor attacks will be greatly reduced. In other words, the backdoor attacks are vulnerable to various transformations, and are sensitive to the difference between the training trigger and the testing trigger. Pasquini and Böhme [22] studied the vulnerability of backdoors in the DNN based face recognition models. They demonstrated that, different geometric and color transformations on the backdoor triggers can significantly restrict the effectiveness of the backdoor attack.

Backdoor attacks in real physical world. To the best of the authors' knowledge, there is only one work conducted on the physical backdoor attacks [13] (an unpublished work in arXiv). Wenger et al. [13] collect 9 different facial accessories in the real world as backdoor triggers, and evaluate the attack effectiveness of these backdoors on the face recognition

models. However, their attacks are carried out under the ideal physical conditions, where the attacker is facing the camera with a proper distance. Moreover, the backdoor triggers used by the attackers in the testing and the training phases are the same (same positions and shapes). In real world attacks, these physical conditions will be greatly restricted, which will seriously degrade the performance of backdoor attacks. In this work, we explore the robustness of backdoors in the physical world for the first time, and propose the PTB method to implement the robust physical backdoor attacks. The proposed method performs various physical transformations on these injected backdoor instances in the training phase, so as to improve the physical robustness of the backdoors. As a result, the attacker can achieve a high attack success rate under these extremely complex and difficult physical conditions.

III. Robust Physical Backdoor Attack Method

In this section, we elaborate the proposed robust physical backdoor attack method, PTB. For the ease of understanding, this paper takes the DNN based face recognition system as an example for discussion, which has been widely deployed in the real physical world.

A. Overview

First, we introduce the overall procedure of the proposed backdoor attack method, which can be divided into the following three steps: 1) generating the backdoor face images; 2) injecting the backdoor instances into the clean training set, and performing the physical transformations on the backdoor instances before each round of training, then, the model is trained to embed the backdoor; 3) triggering the backdoor attacks in the real physical world. Figure 1 presents the overview of the proposed PTB backdoor attack method.

- Generating the backdoor face images. In this paper, we assume that the attacker launches the backdoor attacks against the DNN based face recognition system in real physical world by using data poisoning. Therefore, the attacker requires to generate a batch of backdoor face images (i.e., face images injected with triggers), and inject these backdoor instances into the clean training set to train the target model, so as to embed the specific backdoor. To this end, the attacker collects some face photos of different individuals in the real physical world, where the people all wear the facial accessories (i.e., the injected backdoors). Then, the collected face images will be pre-processed (e.g., cropping and scaling) to ensure that they are close to these face images in the clean training set.

- Injecting backdoor face images & performing physical transformations. When the backdoor face images are generated, the attacker will modifies the labels of these backdoor instances based on the specific target victim to implement the targeted backdoor attacks. Then, these backdoor instances with modified labels are injected into the training set of the target face recognition model, and the model is trained to embed the backdoors. Note that, the goal of proposed attack method is to guarantee that, an injected backdoor can still effectively trigger the attacks after undergoing a variety of physical transformations. For the proposed attack method, at each iteration of the model training, the attacker performs a series of physical transformations T (distance, angle, rotation, lighting, and noise) on all the backdoor instances, so as to enhance the robustness of the injected backdoor. In this way, even under the complex physical conditions, the embedded backdoor can still be successfully triggered, and the proposed backdoor attack can achieve a high performance in real physical world.

- Triggering physical backdoor attacks. In this step, the attacker triggers the backdoor attacks in the physical world. For example, if the target person is t, any attacker using the backdoor would be incorrectly classified as the class t by the attacked face recognition model. In addition to implementing the backdoor attacks under these normal conditions, this paper focus on evaluating the effectiveness of proposed attack method under the complex physical scenes.

B. Physical Transformations

For the backdoor attack against the DNN based face recognition model, the goal is to maximize the probability that a submitted backdoor face image is incorrectly classified as the target class. Meanwhile, the benign input should be classified as the ground-truth label, so as not to be noticed. The goal of backdoor attacker can be formalized as follows:

$$\begin{aligned} &\max\ P_r(F(x+\delta)=t) \\ &\text{s.t. } P_r(F(x)=y) \end{aligned} \tag{1}$$

where x is a clean face image, and $x+\delta$ is the backdoor face image that generated by injecting the trigger δ. y is the ground-truth label of image x, while t is the label of the target class that specified by the attacker. F represents the target face recognition model, and P_r is the confidence of the class that predicted by the model F.

As discussed in Section I, the backdoor attacks are restricted by a variety of physical conditions, which will greatly degrade the attack effectiveness. Inspired by the existing robust physical adversarial example attacks [23], [24], [25], the proposed PTB method aims to simulate these possible physical constraints (that an injected trigger may experienced in the real physical world) in advance to enhance the robustness of backdoor. More specifically, at each iteration of model training, the PTB method performs the physical transformations on these injected backdoor instances, i.e.:

$$\underset{t\in T}{E}[\sum_{i=1}^{m} t(x_i+\delta)] \tag{2}$$

where T is the distribution of physical transformations, and m is the total number of injected backdoor instances.

Note that, to ensure that an attacker can also successfully launch the backdoor attacks under those normal physical conditions, in each round of training, there is a 50% probability of not undergoing transformations (unchanged), as shown in Figure 2. In other words, the proposed PTB method transforms each backdoor face image with a certain probability (50% in this paper) at each iteration.

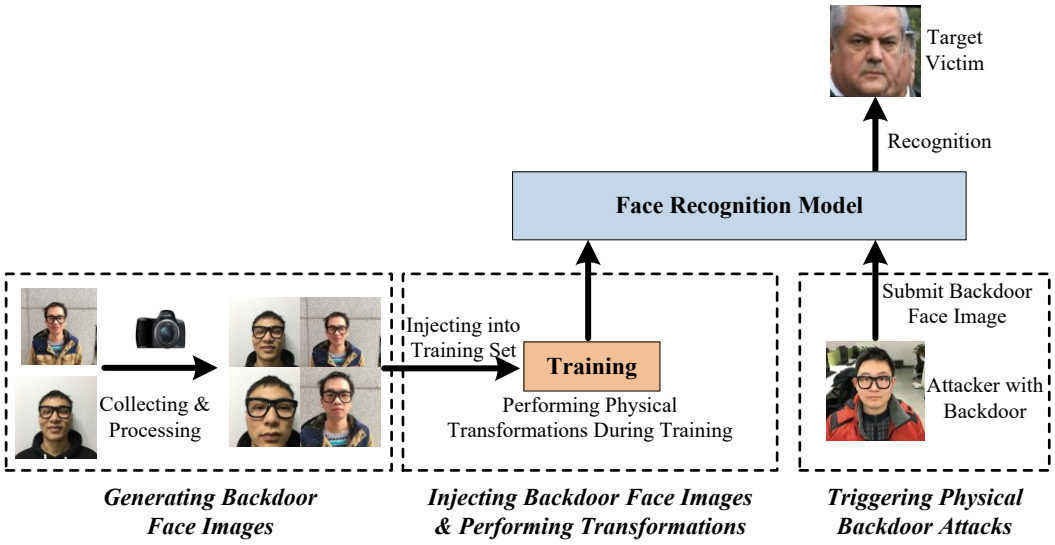

Fig. 1: Overview of the proposed robust physical backdoor attack method.

$$\underset{t\in T, p\in[0,1]}{E}[\sum_{i=1}^{m} p\cdot t(x_i+\delta)+(1-p)\cdot(x_i+\delta)] \quad (3)$$

where p represents the probability of transformation, which takes value of 0 or 1 with 50% probability.

In this way, for the proposed robust physical backdoor attack method PTB, the objective function can be formalized as follows:

$$\max\ P_r\{F[\underset{t\in T p\in[0,1]}{E}(\sum_{i=1}^{m} p\cdot t(x_i+\delta)+(1-p)\cdot(x_i+\delta))]\} \quad (4)$$

In this paper, we consider 5 different physical transformations that a backdoor trigger most likely experienced during the face recognition process in real physical world, i.e., T={Angle, Distance, Rotation, Brightness, Gaussian Noise}. The transformations of the proposed PTB backdoor attack method at each iteration are presented in Figure 2.

- **Distance.** The *Distance* transformation simulates the distance changes between the backdoor trigger and the camera. For the proposed PTB method, the injected backdoor face images will be scaled at random, which ranges from 0.8 to 1.2. In this way, the attacker with the backdoor can effectively trigger the attacks at different distances.
- **Angle.** The *Angle* transformation rotates the backdoor face image at a random angle (horizontal and vertical) that ranges from $0°$ to $90°$. The horizontal rotation (i.e., left-right) simulates the physical constraint in which the camera captures photos at the horizontal angles, while the vertical (i.e., up-down) rotation simulates the physical constraint in which the face images are captured at different vertical angles.

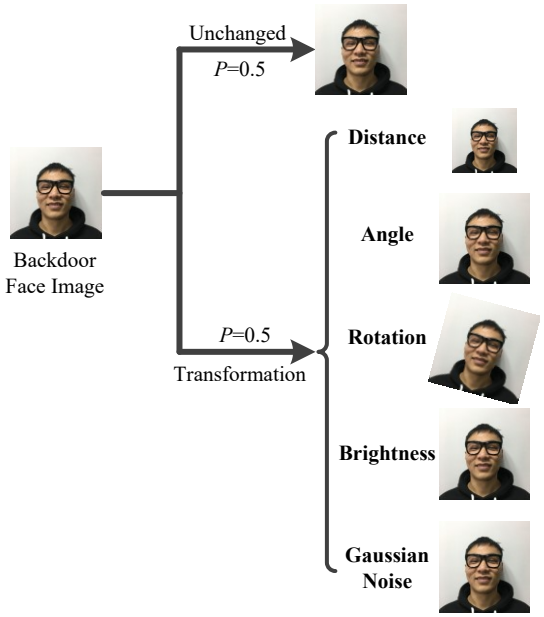

Fig. 2: The physical transformations performed on the injected backdoor face images at each iteration of model training.

- **Rotation.** The *Rotation* transformation simulates the rotation of the backdoor trigger that caused by the swing of an attacker's face. In each iteration, the backdoor face image will be randomly rotated by a certain angle on the 2D plane.
- **Brightness.** To ensure that the attacker can utilize the backdoor to trigger the backdoor attack under different lighting conditions, the proposed PTB method uses the *Brightness* transformation to transform the injected backdoor face images with different brightness, so as to make the trigger adapt to various physical lighting conditions.

- **Gaussian Noise.** The inputs of physical face recognition model are captured by the external camera and preprocessed by the system, which may introduce environment noise on the backdoor. To this end, the *Gaussian Noise* transformation adds the gaussian noises at each injected backdoor face image, which simulates the physical backdoor attacks under the noise condition.

IV. Experiments

A. Experimental Setup

Dataset. In this paper, the experimental evaluations are carried out on a large and realistic dataset, the YouTube Aligned Face dataset [15], which is composed of face images that extracted from the YouTube videos. This dataset contains 1,595 different people, each of which has a different number of face images, ranging from tens to thousands [15]. In our experiment, we crop all the face images and resize them to 224×224. Limited to the size of each image and the scale of the DNN model, we randomly select 100 different persons from the dataset as the experimental data, where each person has 120 face images. 100 images are used for model training, and the remaining 20 images are used for testing. In this way, our experimental dataset contains a total of 12,000 different face images.

DNN models. The target DNN model of the proposed backdoor attacks is VGGFace [14].

- VGGFace [14]. The VGGFace is a 16-layer standard face recognition model, which consists of 13 convolutional layers and 3 fully connected layers. We adopt the experimental settings in these exisiting works [6], [13] and download the VGGFace model that pre-trained on the ImageNet [26]. The last three layers of the VGGFace model are fine-turned, and the sofatmax activation layer is replaced, so as to train the VGGFace model on our experimental dataset. The model is trained on the YouTube Aligned dataset for 30 epoches, and the batch size is set to be 64. In our experiment, the test accuracy of VGGFace model achieves 96.33% without the backdoor attacks.

Metrics. The proposed PTB method is evaluated with the following two metrics.

- Success rate of backdoor attacks R_{ptb}. This metric indicates the proportion of backsoor face images that are classified as the target class t among all submitted face images, which is computed as follows:

$$R_{ptb} = \frac{F_t}{F_s} \times 100\% \tag{5}$$

where F_t represents the number of face images that classified as the target calss t, and F_s represents the number of all the submitted backdoor images.

- Performance drop of target DNN model D_{ptb}. This metric represents the degradation of target model's test accuracy that caused by the proposed backdoor attack, which is calculated as follows:

$$D_{ptb} = ||D_c - D_b|| \tag{6}$$

where D_c represents the test accuracy of target model that trained on the clean face images, while D_b represents the test accuracy of the DNN model that trained on the backdoored training set.

B. Experimental Results

In this section, we evaluate the performance of the proposed backdoor attack method PTB on the target face recognition model. Specifically, we implement two different types of backdoor attacks. First, without the PTB method, we directly inject these backdoor face images into the clean training set, and train the target face recognition model to embed the backdoor. Second, with the PTB method, a series of physical transformations will be performed on these injected backdoor face images during each round of training process. In addition, the backdoor attacks are implemented in two attack scenarios: (i) *"Simple Scene"*, represents the normal recognition scenario, in which the backdoor sample is not restricted by physical conditions during the attack, i.e., the attacks are launched under an ideal physical condition. (ii) *"Complex Scene"*, represents the complex recognition scenario, in which the backdoor instance suffers from a variety of complex physical conditions during the attack.

In our experiments, to evaluate the effectiveness of the proposed PTB method, we have collected the attacker's face images that injected with the backdoor under the above *Simple* and *Complex* attack scene, respectively. For each experimental setting, the backdoor embedding and attacking process are repeated for 5 times, and each time a specific target victim is specified. 20 photos are taken for each setting. As a result, 5 attack results are obtained for each setting, and we reported the minimum, maximum, and average of these 5 results in this paper.

1) Attack Effectiveness on VGGFace Model: First, we evaluate the proposed backdoor attack method on the VGGFace model [14]. To launch the attacks, we exploit a black square with the size of 6cm×6cm as the backdoor trigger (referred as *Square*). In the real physical world, we paste the *Square* backdoor to the forehead of the attacker, and use a camera to capture the photos under the *Complex* and *Simple* scenes, respectively. Then, these backdoor images will be submitted to the target face recognition model to evaluate the effectiveness of the proposed method. The attack results are presented in Table I.

TABLE I: Physical attack performance of backdoor *Square*

Scenes	Simple			Complex		
	min	max	**ave**	min	max	**ave**
Without PTB	75%	100%	**91%**	0%	25%	**5%**
With PTB	95%	100%	**99%**	65%	90%	**78%**

It is shown in Table I that, without the PTB method, the performance of backdoor attacks under the *Simple* scene is high, where the maximum and average attack success rate is 100% and 75%, respectively. However, the success rate of backdoor attacks without the PTB method drops sharply under the *Complex* scene, where the attack success rate is

only 0% (minimum), 25% (maximum) and 5% (average), respectively. This indicates that the backdoor attacks without the PTB method almost completely failed under the complex physical conditions. With the proposed PTB method, the performances of backdoor attacks are much better than the backdoor attacks without the PTB method under both *Simple* and *Complex* conditions. Under the *Simple* scene, the attack success rate of backdoor attack with PTB is 95%, 100%, and 99%, respectively. Under the *Complex* scene, the average attack success rate has reached 78%, while the highest attack success rate is high up to 90%. After using the proposed PTB method, the performance of backdoor attacks under the complex physical conditions has increased from 5% to 78%, which demonstrates the effectiveness of our proposed method.

2) Attack Performance of Different Triggers: In this paper, the proposed PTB method is effective and robust for different types of backdoors. To demonstrate this, in addition to the backdoor *Square*, this paper also evaluates the effectiveness of the proposed PTB method with other types of backdoors. Specifically, a black triangle backdoor (referred as *Triangle*) and a pair of black-frame glasses (referred as *Glasses*) are exploited as the trigger of backdoor attacks, respectively. Similarly, we paste these two backdoors on the face of the attacker, and then take photos to perform the physical backdoor attacks.

TABLE II: Attack performance of two different backdoors in the physical world

Backdoor	**Scenes**	**Simple**			**Complex**		
		min	max	**ave**	min	max	**ave**
Triangle	Without PTB	85%	100%	**96%**	0%	35%	**11%**
	With PTB	85%	100%	**97%**	60%	95%	**82%**
Glasses	Without PTB	90%	100%	**96%**	0%	20%	**9%**
	With PTB	100%	100%	**100%**	70%	90%	**79%**

The attack performance of the two different backdoors in the real physical world are presented in Table II. It is shown that, without the proposed PTB method, the backdoor attacks perform well under the *Simple* scene, where the average attack success rate of two backdoors (*Triangle* and *Glasses*) both achieves 96%. However, under the *Complex* scene, the attack performance of these two backdoors are rather poor. The average attack success rate is only 11% (*Triangle*) and 9% (*Glasses*), respectively. Once the PTB method has been exploited during the training process of target model, the performance of these two backdoors are greatly improved. The average attack success rate under the *Complex* scene reaches 82% and 79%, respectively. Meanwhile, the attack performance under these simple scenes are also high.

3) Attack Performance of Different Positions: In the above experiments, the backdoors (*Square*, *Triangle* and *Glasses*) are pasted on the forehead area of a human face. In fact, for a backdoor that pasted on other different positions, the proposed PTB method is also feasible and effective. Specifically, we paste a black square backdoor with the size of 6cm$\times$6cm on the chin of a human face, so as to evaluate the impacts of backdoors pasted on different positions on the proposed attack method. The experimental results are shown in Table III. It is shown that, when the backdoor pasted on the chin of a human face, the proposed PTB method is still effective. The average attack success rate of backdoor *Square* has improved from 10% (without PTB) to 75% (with PTB) under the complex physical conditions. Meanwhile, the performance of backdoor attacks under the *Simple* scene is also higher than that without PTB.

TABLE III: Performance of proposed PTB method when the backdoor pasted on different position (i.e., on the chin of a human face)

Scenes	Simple			Complex		
	min	max	**ave**	min	max	**ave**
Without PTB	85%	90%	**87%**	0%	30%	**10%**
With PTB	90%	100%	**97%**	65%	85%	**75%**

4) Impacts of Different Injection Number: Finally, we explore the impacts of different number of injected backdoor face images on the proposed PTB method. Specifically, we select the person with the label of "022" in the YouTube Aligned Face dataset [15] as the target victim, and use the backdoor *Square* to implement backdoor attacks. The backdoor face image generation and the detailed attack process are the same as that described in Section IV-B1. In this experiments, the number of backdoor face images that injected into the training set is 5, 10, 20, 50, and 100, respectively.

The attack success rate under different numbers of injected backdoor instances is shown in Figure 3. It is shown that, under the *Simple* scene, the performance of backdoor attack with and without the proposed PTB method is close. After injecting 50 backdoor face images, the success rate of the backdoor attack reached the upper bound of attack performance (*i.e.*, 100%). However, under the *Complex* scene, the performance of the backdoor attack with the PTB method is much better than the performance of the backdoor attack without the PTB method. Specifically, after injecting 50 backdoor face images, the success rate of the backdoor attack with the PTB method has reached 90%, while the success rate of the attack without the PTB method is only 5%. The results indicate that: (i) As the injection number increases, the performance of backdoor attacks with the PTB method improves. (ii) When the injection number reaches 50, which only accounts for 0.5% (50/10000), the performance of the backdoor attack has reached a very high value. This indicates that, by injecting a very small ratio (0.5%) of backdoor instances, a high attack success rate can be achieved. (iii) Under the *Complex* physical conditions, the performance of the backdoor attack without the PTB method will be restricted even injecting with more number of backdoor instances, while the backdoor attack with the PTB method is robust.

V. Conclusion

This paper studies the robustness of backdoor attacks in real physical world. By performing various physical transformations during each iteration of model training, the proposed PTB attack method greatly guarantees the physical robustness and effectiveness of backdoors. The transformations simulate

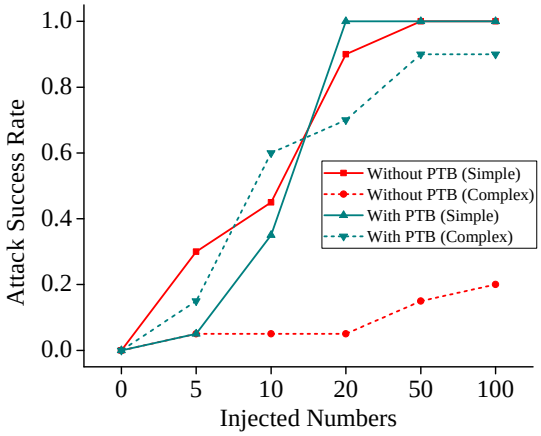

Fig. 3: The physical transformations performed on the injected backdoor face images at each iteration of model training.

these physical constraints that a backdoor may experience in real physical world, which improves its robustness under real complex physical conditions. To the best of the authors' knowledge, this paper is the first research on the robustness of backdoor attacks in physical world. Experimental results demonstrate that, the proposed PTB method can significantly improve the attack success rate of backdoor attacks on the state-of-the-art face recognition model (VGGFace), especially under those complex physical scenes. Meanwhile, the normal performance of target recognition models are not affected, thus the proposed backdoor attack is convert and is hard to notice. In the future work, we will explore the effectiveness of the proposed physical backdoor attack method on other DNN classifiers and the more complex object detectors.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (no. 61602241).

References

[1] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales, "When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 142–150.

[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.

[3] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[4] M. Bojarski, D. D. Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," *arXiv:1604.07316*, 2016.

[5] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[6] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv:1712.05526*, 2017.

[7] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden trigger backdoor attacks," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 11 957–11 965.

[8] Y. Yao, H. Li, H. Zheng, and B. Y. Zhao, "Latent backdoor attacks on deep neural networks," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 2041–2055.

[9] A. S. Rakin, Z. He, and D. Fan, "TBT: targeted neural network attack with bit trojan," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 195–13 204.

[10] C. Guo, R. Wu, and K. Q. Weinberger, "Trojannet: Embedding hidden trojan horse models in neural networks," *arXiv:2002.10078*, 2020.

[11] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proceedings of the 25th Annual Network and Distributed System Security Symposium*, 2018.

[12] H. Zhong, C. Liao, A. C. Squicciarini, S. Zhu, and D. J. Miller, "Backdoor embedding in convolutional neural network models via invisible perturbation," in *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 2020, pp. 97–108.

[13] E. Wenger, J. Passananti, Y. Yao, H. Zheng, and B. Y. Zhao, "Backdoor attacks on facial recognition in the physical world," *arXiv:2006.14580*, 2020.

[14] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proceedings of the British Machine Vision Conference 2015*, 2015, pp. 1–12.

[15] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 529–534.

[16] M. Zou, Y. Shi, C. Wang, F. Li, W. Song, and Y. Wang, "Potrojan: powerful neural-level trojan designs in deep learning models," *arXiv: 1802.03043*, 2018.

[17] J. Dumford and W. J. Scheirer, "Backdooring convolutional neural networks via targeted weight perturbations," in *IEEE International Joint Conference on Biometrics*, 2020, pp. 1–9.

[18] S. Garg, A. Kumar, V. Goel, and Y. Liang, "Can adversarial weight perturbations inject neural backdoors?" in *The 29th ACM International Conference on Information and Knowledge Management*, 2020, pp. 2029–2032.

[19] S. Li, M. Xue, B. Zhao, H. Zhu, and X. Zhang, "Invisible backdoor attacks on deep neural networks via steganography and regularization," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–12, 2020.

[20] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Proceedings of the 16th European Conference on Computer Vision*, 2020, pp. 182–199.

[21] Y. Li, T. Zhai, B. Wu, Y. Jiang, Z. Li, and S. Xia, "Rethinking the trigger of backdoor attack," *arXiv: 2004.04692*, vol. abs/2004.04692, 2020.

[22] C. Pasquini and R. Böhme, "Trembling triggers: exploring the sensitivity of backdoors in dnn-based face recognition," *EURASIP J. Inf. Secur.*, vol. 2020, pp. 1–12, 2020.

[23] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 284–293.

[24] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[25] S. Chen, C. Cornelius, J. Martin, and D. H. P. Chau, "Shapeshifter: Robust physical adversarial attack on faster R-CNN object detector," in *Proceedings of Machine Learning and Knowledge Discovery in Databases*, 2018, pp. 52–68.

[26] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.